A DELPHI ASSESSMENT OF THE DIGITAL ROSETTA STONE MODEL

THESIS

Don M. Kelley, Captain, USAF

AFIT/GIR/ENV/01M-10

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GIR/ENV/01M-10

A DELPHI ASSESSMENT OF THE DIGITAL ROSETTA STONE MODEL

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Information Resource Management

Don M. Kelley, B.S.

Captain, USAF

March 2001

AFIT/GIR/ENV/01M-10

A DELPHI ASSESSMENT OF THE
DIGITAL ROSETTA STONE MODEL

Don M. Kelley, B.S.
Captain, USAF

Approved:

Alan R. Heminger — 5 Mar 2001

Alan R. Heminger (Chairman) — date

David P. Biros — 5 Mar 2001

Maj. David P. Biros (Member) — date

Mark A. Ward — 5 Mar 2001

Maj. Mark A. Ward (Member) — date

Acknowledgments

I would like to express my sincere appreciation to my wife for her continued support in this exciting but challenging work. Without you, dear, none of this would have been worthwhile or even possible. And to our little one—as yet unborn—this has been an exciting and wonderful time anticipating your arrival. I also need to thank my parents for their support and encouragement. You always made me feel I could accomplish anything I set my mind to. And to my cat, I owe a debt of gratitude for supervising my late-night bouts of typing. He had a keen sense of knowing when the computer had had enough attention and when it was time for him to step in.

I am also indebted to Dr. Alan Heminger, Major Ward and Major Biros. Thanks for your perseverance through this "It's the thesis that never ends" process. Your guidance and support is much appreciated. I would be remiss if I didn't mention Albert, Richard, and Orval. Thanks guys, you really made my experience here something to remember.

And the most important and well-deserved thanks go to Jesus Christ, my personal saviour, Lord, and true friend. I'm continually amazed by your infinite wisdom and your never-ending desire for an intimate relationship with me. If there is truth anywhere in this thesis, it is this, what you have spoken in John 14:6—"I am the way and the truth and the life. No one comes to the Father except through me."

Don M. Kelley

Table of Contents

Page

Page

List Of Figures

Page

List of Tables

Page

AFIT/GIR/ENV/01M-10

Abstract

Information that is stored digitally can only be used if it can be retrieved and interpreted. If the methods to retrieve the information are lost, it may be difficult, if not impossible to re-create them. The knowledge to interpret the bitstream is also at risk. The Digital Rosetta Stone (DRS) Model was developed as a framework for capturing and maintaining the methods necessary to retrieve and display digital information stored on obsolete media or using obsolete software. However, this conceptual model had not yet been assessed by the community of professionals for its practical efficacy. This thesis began the assessment process by using the Delphi Method to explore the DRS with those responsible for maintaining access to digital data.

The literature review found several strategies for maintaining long-term access, but also found them to be mostly preservation oriented. These strategies sometimes address recovering the information, but require some prior action at storage time that would allow for recovery later. The DRS model is designed for those situations where no specific preservation strategy was employed or is unknown, so there is nothing to work with except the stored bitstream of the document. The DRS focuses on recovering both the bitstream and interpreting that bitstream to re-create the original document.

During the first round of the Delphi, the ideas expressed by the group of experts formed the basis for further discussion. Overall, the group expressed concerns about the practicality of developing the DRS, but agreed that it is an important concept that should be explored further. If found to be technologically feasible and economically desirable, the DRS could well lead to a long-term solution for recovering information that would otherwise be impossible to recover.

A DELPHI ASSESSMENT OF THE
DIGITAL ROSETTA STONE MODEL

I. Introduction

Background

This study deals with accessing information in computing devices that is many generations behind the current technology. People have been storing information since the dawn of human history. However, only relatively lately have people begun to store information in digital format. What makes this important is that for the first time we are beginning to store much of our historically important information in a way that cannot be read without specific, often esoteric technologies that we may well lose.

Digital Storage: A Problem That Has Been A Long Time Coming

As we have gone through the years using a series of technologies for storing information digitally, we have amassed a tremendous amount of information. Academic institutions, libraries, and other digital material storehouses, such as museums, have some sort of digital archive. The Massachusetts Institute of Technology (MIT), for example, has rooms full of magnetic tapes, some dating back to the early 1970's (Zuzga, 1995). A recent survey by Hedstrom and Montgomery (1998) found that of 19 such digital storehouses, there was a total of at least 4.1 terabytes.

A Growing Problem

Today, more information is being stored digitally than was thought possible even a few years ago. World production of unique information has been estimated to be between one and two billion gigabytes (one to two exabytes) per year (Lyman and Varian, 2000). This table shows the scale of the binary powers of 10.

Binary Powers of Ten

- Bit = 1 decision (the smallest unit of storage)
- Byte = 8 bits
- Kilobyte = 1,024 or 2^{10} bytes
- Megabyte = 1,048,576 or 2^{20} bytes
- Gigabyte = 1,073,741,824 or 2^{30} bytes
- Terabyte = 1,099,511,627,776 or 2^{40} bytes (~1 Thousand GB)
- Petabyte = 1,125,899,906,842,620 or 2^{50} bytes (~1 Million GB)
- Exabyte = 1,152,921,504,606,850,000 or 2^{60} bytes (~1 Billion GB)

One terabyte is the equivalent of printing about 50,000 trees worth of paper (Lyman and Varian, 2000). One petabyte, in terms of storage requirements, is about half of all United States Academic Research Libraries (ibid.).

Militaries, governments, and private groups are storing thousands of gigabytes every day (Williams, 2000)—perhaps hundreds of thousands. Over the last 25 years, the National Archives and Records Administration (NARA) has received approximately 90,000 electronic records, which was just a fraction of what was produced by the entire U.S. Government. The U.S. Treasury Department alone is now generating some 960,000 electronic-mail files annually (Carlin, 1998).

NARA was created as a repository for government documents and other historically significant materials. In accordance with 44 USC § 3102, the head of each Federal agency is charged with cooperating with NARA in the "selection and utilization of

equipment and supplies associated with records." The Archivist of the United States, in turn, is required to accept "sufficiently historical or otherwise valuable records" (44 USC § 2107). However, "NARA faces increasingly enormous quantities of records" (Carlin, 1998). As if shear volume was not enough of a problem, NARA is also receiving "an increasingly diverse load of [digital] information" created using a wide variety of software and stored in a "bewildering variety of media" (Smith, 1998:4). This predisposes information to the threat of being permanently lost, even if it is under NARA's watchful eye.

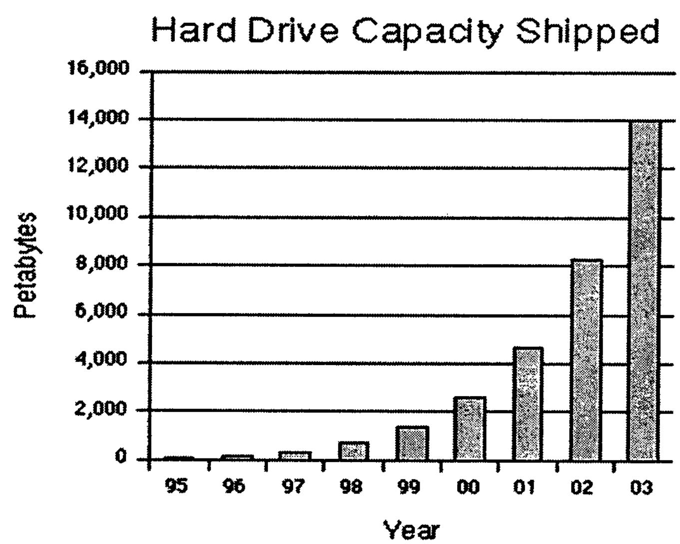


Table 1: 1999 Winchester Disk Drive Market Forecast and Review

Six years ago, in 1995, the hard drive industry shipped about 104.8 petabytes of storage (Lyman and Varian, 2000), and the magnetic tape industry shipped about 200 petabytes of storage (Lesk, 2000; Williams, 2000). Since 1995, over 5.3 exabytes of hard drive storage have been shipped. "Industry rules of thumb suggest that there is about 10

times as much storage on tape as on hard drives" (Lyman and Varian, 2000:5). This growth shows that the technological capability exists to store huge amounts of data.

Threats To Our Digital History

Digital preservation is, in some ways, similar to traditional information preservation. Some of the similarities are that they share methods for identifying, cataloging, and physically storing the material. However, digital preservation is a relatively new field (Smith, 1998); and we have not yet developed effective countermeasures to some of the threats our digital history faces. "A 'digital gap' will span from the beginning of the wide-spread use of the computer until the time we eventually solve the problem" (Carlin, 1998:3). Our digital history includes everything stored digitally--our government entitlements, public business documents, and myriad other records (ibid.).

It appears that the amount of data that we create is growing exponentially (Carlin, 1998; Moore, *et al.*, 2000; Lyman and Varian, 2000). In Table 1, the exponential growth curve is evident. The sheer magnitude of new data that are being added to the already large store of digital information exacerbates the problem of managing it all. The Archivist of the United States put it eloquently when he said, "It will be worse than sad if the marvelous technologies that are giving us a new information age outrun our ability to keep a record of it" (Carlin, 1998). Federal agencies are already facing the problem of "long-term storage and access of digital information" (Moore, *et al.*, 2000:1).

Data Creators/Owners/Providers. One may think that the creators and handlers of the data would act appropriately to ensure the viability of their data. However, they can threaten access to the data by failing to accept the responsibility of preserving it or ensuring its preservation (TFADI, 1996). The creators may only want the data for a short while, not recognizing that it may have some historical importance (Zuzga, 1995). They may also not adopt behaviors that will facilitate the preservation of the data (Beagrie, 1998). One of the ways to facilitate preservation of the data is to make sure that it resides on currently accessible storage devices and can be accessed by current software. In other words, these people could strive to stay ahead of technological obsolescence. However, as we continue to accumulate and store information digitally, we are becoming overwhelmed by any efforts to ensure that all potentially valuable information is still accessible in its original technology or that it is moved to a newer, still supported technology.

Technological Obsolescence. A technological generation, for purposes of this paper, is loosely defined as a time period in which the hardware and/or software perform in some similar range. A computer related example is the difference between 5¼ inch and 3½ inch floppy disks. They can store the same types of data, but need different devices for access. They also store the same types of data, but in different amounts. An example not from the computer field is the generation gap between 33 1/3 rpm records and reel-to-reel tape. The differences between a record and reel-to-reel tape are quite large—media incompatibility, data rates, configuration of devices are significant differences. They also require different technologies to access the information stored on

them. One typically realizes that equipment is technologically obsolete when a majority of the population is using hardware and software of a newer technological generation. Another indicator of technological obsolescence is a diminishing support base.

Over the normal lifetime of a computer, maintenance issues are expected. The longer it is used, the more it costs to maintain it, for two reasons. First, the number of people who know how to fix the problem shrinks. Second, there are fewer and fewer spare parts. The high cost of maintaining obsolete technologies has "hindered the preservation of film and video materials" (MacCarn, 2000:1). NARA has already stumbled across technological obsolescence more than a few times. The Sony Model 800 tape machine, Nagra TRVR recorder, Dictabelt machine, and the Zapruder family camera that filmed John F. Kennedy's assassination are all obsolete and very difficult to find in working condition (Carlin, 1998) – and without these machines, the ability to access information stored in these formats may be lost forever.

Traditionally, the rate of technological change has advanced slowly. With this slow transition, "both old and new versions of the software and hardware infrastructure [were] present at the same time" (Moore, *et al.*, 2000:7). However, the explosive growth of the Internet and a closely related flurry of e-business activity have created a sort of ever-increasing Technological Arms Race.

Accelerating Obsolescence. This "race", created by accelerating technological development also creates accelerating technological obsolescence. It is this obsolescence which is threatening our knowledge of methods used to retrieve and properly display our stored digital history (Smith, 1998; Graham, 1997; Lyman and Besser, 2000; TFADI,

1996; Day, 1997; Carlin, 1998; Kenney, 1996). Accelerating obsolescence is the quickly shortening lifespan of widely used storage and computing devices (including both hardware and software) (Kenney, 1996). The lifespan shortening occurs simply because the time between technological generations is becoming shorter. The impetus for this race occurs when companies (both producing and consuming), get caught up in a paradigm of needing to push the technological envelope in order to stay competitive. Hodge (1999), Moore, *et al.* (2000), Lyman and Varian (2000) and many others see an accelerated rate of increase in this race. This increase speeds up what I call a "revolving door of technologies".

Hedstrom (2000) suggests that accelerating technological obsolescence poses the greatest danger to our digital history. Even if a decision is made to move all stored information to a new storage technological generation, we may not be able to move all of the data to a current form in a current storage environment before the hardware- and software-technological gap becomes too wide. The time it takes to transfer the data to the new generation may be longer than the life of the new generation. The question then becomes, "Should I finish this project as is, or should I try to move information from both older generations to the new one?" As one systems librarian puts it, "All those state-of-the-art machines, software packages, and compression techniques seem old before the boxes and shrink-wrap even hit the landfill" (Pace, 2000:55). Regardless of the chosen solution, the problem is compounded by an exponentially growing data set. This natural progression of accelerating obsolescence can be seen when Moore's Law is considered.

Moore's Law. This unscientific law was originally an observation regarding circuit capacity on computer chips. It has taken on much broader implications than the computing field, but directly correlates to the length of technological generations.

> In 1965, Gordon Moore was preparing a speech and made a memorable observation. When he started to graph data about the growth in memory chip performance, he realized there was a striking trend. Each new chip contained roughly twice as much capacity as its predecessor, and each chip was released within 18-24 months of the previous chip. If this trend continued, he reasoned, computing power would rise exponentially over relatively brief periods of time (Intel, 2000b).

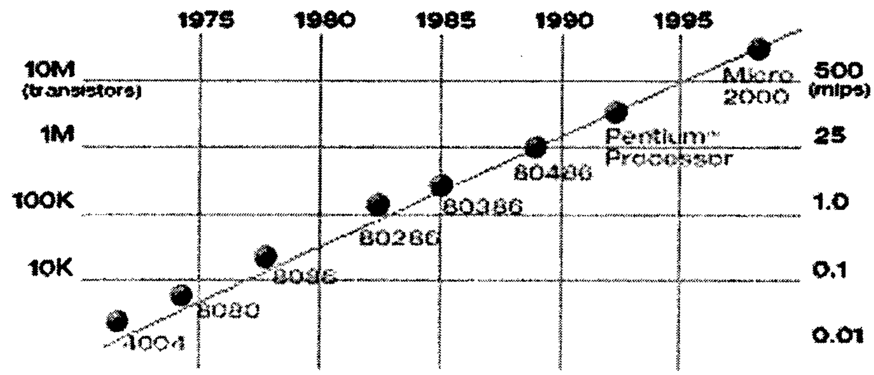


Figure 1: History of the Microprocessor (Intel, 2000a)

Although there is some conjecture that Moore's Law is self-fulfilling, it is none-the-less widely used as a technological barometer (Schaller, 1996).

Our love affair with cutting edge technology introduces an interesting paradox. We become so consumed with having the latest and greatest technology that we cast aside today's technology for tomorrow's technology without considering the consequences. In other words, we have become addicted to speed (Schaller, 1996). There would be no problem with this addiction, except that in our haste to have the best, we sometimes burn our bridges to the past. We even do this gladly if we think it will help us move on. When Cortés, the Spanish conquistador, faced pleas to return home, he burned his ships. As a

result, his troops were well motivated to survive where they were (Encyclopedia Britannica, 2001). It is not always wise to take this approach because some consequences are overlooked. One of these consequences is known as media degradation.

Media Degradation. Everything wears out over time. Some things wear out sooner than others. Storage media are not exempt from physical decay, extended use, and other factors that degrade the information stored on them, such as abuse and neglect. This wear and tear on media is known as media degradation. "The default condition of paper is persistence, if not interrupted; the default condition of electronic signals is interruption, if not periodically renewed" (Lyman and Besser, 2000:7). However, as serious as media degradation is to our stored information, technological obsolescence is far more a threat to our digital history than media degradation because it tends to occur much faster than media degrade (Zuzga, 1995; Graham, 1997; TFADI, 1996). This thesis does not address resolving media degradation problems.

Some Data Will Be Left Behind. Limited resources preclude efforts to preserve all available information objects (TFADI, 1996). Tennant (1998) suggests that one criterion to consider when maintaining long-term access to digital documents is determining how much material is enough to make access efforts worthwhile. If this were the only criterion used to determine which information is preserved, small disparate amounts of data could be stranded.

Another criterion is the enduring value of the information (National Archives of Canada, 1995). Document retention schedules set up by the U.S. Government recognize the value of maintaining legal documents for long periods of time. However, documents

deemed less significant are usually destroyed or kept for only a short while. The criteria for determining document value can change depending on many and varied factors such as world events or high-profile lawsuits. This can lead to information being lost that is later deemed important but unrecoverable. In trying to upgrade some of MIT's archival data, Brian Zuzga (1995) realized that there were some tapes that, although not currently valuable, may be of extraordinary value in the future.

The sad reality is that the overlooked consequences are already cropping up—we have irretrievably lost critical data on more than a few occasions (Robertson, 1996; TFADI 1996). Due to negligence, mishandling, and technological obsolescence, most of Canada's early recordings of feature films, radio broadcasts, and video are forever lost (National Archives of Canada, 1995). While the content of the first electronic mail message may be remembered, the actual message, sent in 1964, has not survived (TFADI, 1996). Accelerating technological obsolescence has struck Professor Hans Rollman who kept some important data on eight inch floppy disks primarily used in the 1980s. In a posting on the Internet, he pleads for anyone who can help him access his "imprisoned data" (Poitras, 1998).

A Solution To The Threats. The key to knowledge preservation is to capture the information about storage devices and software algorithms when it is readily available. Generally, this means while the storage device and software are still in general use. It is the key to knowledge preservation because if one waits until the technology is no longer current, it may be too late. For example, when trying to recover the information stored on an 8-Track Punched Paper Tape, Robertson (1996) was not able to map out the entire

storage technique, even though the technology was relatively recent and the company that designed the technique was still a current industry leader. If the relatively simple task of being able to read paper tape is difficult, then the task of reading magnetic media without the proper know-how is truly daunting. Now imagine the problem if it occurs 100 years later.

Problem To Be Addressed By This Research

Just as the original Rosetta Stone was used to unlock the mysteries of stored written information, the Digital Rosetta Stone (DRS) Model was developed to recover the digital bitstream from an obsolete medium and interpret that bitstream so the information can be properly displayed (Robertson, 1996). However, it is still a conceptual model. It is a framework of ideas that could be a solution to the long-term access problem. This thesis seeks to present the DRS to representatives of the preservation and access community and to build a common understanding about the model based on expert opinion. This opinion will be elicited via the Delphi Method.

Research Question. The research question to be explored in this Delphi Study is, "Is the DRS model a potentially useful method for maintaining long-term access to digital documents?". The following sub-questions were developed to answer the research question.

1. What are the strengths of the Digital Rosetta Stone Model?
2. What are the areas in the Digital Rosetta Stone Model that need improvement?
3. What is missing from the Digital Rosetta Stone Model?
4. How does the Digital Rosetta Stone Model compare with other models in relation to maintaining long-term access to digital documents?
5. What are the underlying assumptions of the Digital Rosetta Stone Model?

6. What steps are necessary to begin implementation of the Digital Rosetta Stone Model?
7. Who should undertake development and implementation of the Digital Rosetta Stone? And why?
8. Do the experts have anything else to contribute that does not fit in the previous questions?

Preview

The next chapter contains the literature review that covers what is known and published on this topic. Chapter III discusses the methodology of research for this thesis. Chapter IV reports findings and limitations of the study. Finally, Chapter V presents a discussion of the findings and recommendations for further research.

II. Literature Review

Chapter Overview

This literature review was conducted to survey what is currently known about the topic of maintaining long-term access to digital objects. The danger of losing our digital history posed by the threats discussed in Chapter 1 is looming on the horizon. To counter this, a number of people have proposed strategies (TFADI, 1996; Willis, 1992; Rothenberg, 1998; MacCarn, 2000). NARA's and other's efforts to date have been labor intensive and expensive. There is also a lack of agreement in the community as to which is the best way to proceed (Kochtanek and Hein, 1999). As will be discussed, each of these strategies offers possibilities, but each also suffers from a number of drawbacks. While some data will be unrecoverable no matter what, the DRS is designed to recover information that would otherwise be left behind by these strategies and when there are no other means of accessing the information. The following sections cover topics that relate to maintaining long-term access and lead into a discussion of these strategies.

Factors Germane To Long-Term Access

Digital Information Objects. Many different types of information can be stored digitally. "Textual, numeric, image, video, sound, multimedia, simulation and so on" are all instances of these many different types (TFADI, 1996:12). We call these stored files digital information objects for purposes of this study. While contextual information can also be gleaned from a collection of files—18 minutes of silence on the Nixon tapes is

more important than just 18 minutes of silence—the DRS is focused on recovering the minutes of silence instead of its significance.

These digital objects can go through several processes during their existence. "Creation, acquisition, cataloging/identification, storage, preservation and access…" are some of those stages (Hodge, 2000:2). These digital objects can also undergo cyclical changes. The rate of change determines whether an object is considered static or dynamic.

Static Versus Dynamic Objects. Static digital objects are those that do not change over time or only have minor changes. For example, a static digital object may be an electronic picture of the Declaration of Independence. A dynamic digital object, however, is one that changes over time. For instance, the Internet web page for CNN.com is dynamic because it changes every few minutes. Information contained in one website on the internet does not necessarily reside all in one place, nor all in one digital object.

Local Versus Non-Local Information. Information that is stored in more than one digital object is non local. If one intended to perform backup procedures on a website, certain questions arise. Should every version of the website be archived? Should the links be archived as well? If so, should the objects of links be archived? A recent government initiative to capture a snapshot of the federal government's web presence quickly revealed how problematic such an endeavor can be (Matthews, 2001; Daukantas, 2001). This research does not attempt to recover digital objects that require references to such non-local information.

Metadata

Metadata can have different meanings to different people (McKemmish, 1998). Simply put, metadata is data about data. "The imprecision of this definition has since allowed it to be applied to any computer-related descriptive information" (ibid.). There are different kinds of metadata, such as recordkeeping, systems operating, data management, information management, discovery, and retrieval metadata (ibid.). Metadata, here, is used for describing digital objects. It is used primarily as part of a key strategy for preserving digital information objects (Day, 1997; Pace, 2000). However, as necessary as metadata is, it does pose certain problems.

Metadata Problems. "The first, and by far the hardest, is a question of what the metadata elements should be" (Miller, 1996:3). Another is metadata management (ibid.). Metadata elements are the different descriptors of the data. Determining what elements should be used in metadata is difficult because it forces us to make forecasts about a future computing environment that is not yet known. Also, there is a related problem of "content standards" or what to put in the metadata elements (Hodge, 2000:6). As a result, there are serious issues that need to be resolved if any metadata initiative is to be successful. Developing a flexible metadata framework can help with this uncertainty.

Dublin Core. One metadata initiative is the Dublin Core. It "is a metadata element set intended to facilitate discovery of electronic resources" (Dublin Core Metadata Initiative, 2000:1). There are ten attributes for element description (Dublin Core Metadata Element Set, 1999:1).

The attributes are:

1. Name
2. Identifier
3. Version
4. Registration Authority
5. Language
6. Definition
7. Obligation
8. Datatype
9. Maximum Occurrence
10. Comment

The above attributes apply to each of the following fifteen elements (Dublin Core Metadata Element Set, 1999:1).

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

The Dublin Core may be useful serving as a "minimal metadata set in order to allow for interoperability between other, more complex metadata formats" (Day, 1997:3). One of those more complex metadata formats is the eXtensible Markup Language.

eXtensible Markup Language (XML). XML is a language that is used on the World Wide Web to provide a universal format and structure for digital objects (Tait, 2000). It is designed more to describe the type of information contained in the document, rather than the actual content (ibid.). One of the key benefits of XML is that it can be tailored to suit any organization's specialized needs by being extensible. However, this flexibility is a two-edged sword. Organizations can tailor XML to suit their individual needs, but this tailoring may not be captured in external documentation. XML facilitates "an infrastructure independent representation" of data (Moore, *et al.,* 2000:3). While they describe the data well, Dublin Core and XML do not capture all of the information relevant to digital objects. Metaknowledge makes up for that shortcoming.

Metaknowledge And How It Is Different From Metadata

Metaknowledge is what we know about how digital information objects are stored, accessed and interpreted (Robertson, 1996). This type of information can represent standards and proprietary algorithms. Hardware and software producers use metaknowledge to build products that can interoperate with each other. A collection of metaknowledge can be used to describe a particular hardware and software environment.

Metaknowledge Is Different From Metadata. Whereas metadata is information about the data, metaknowledge is information about how the data is stored, accessed, and formatted. One example is that metadata is analogous to a library card catalog of information about a book. Metaknowledge, then is knowing that the book has information stored in it and it reads from left to right and top to bottom. Also, metaknowledge is knowing that the black markings on the paper are letters, that when combined, form words that convey written information. The metaknowledge, then for the computing environment, would include everything necessary to retrieve the bitstream from the medium and then properly interpret it.

Criteria for an Ideal Solution

It is clear that a solution to the threats faced by our digital history must meet a certain set of criteria that will ensure its viability. As Rothenberg (1998) puts it

> An ideal approach should provide a single, extensible, long-term solution that can be designed once and for all and applied uniformly, automatically, and in synchrony (for example, at every refresh cycle) to all types of documents and all media, with minimal human intervention. It should provide maximum leverage, in the sense that implementing it for any document type should make it usable for all document types. It should facilitate document management (cataloging, deaccessioning, and so forth) by associating human-readable labeling information and metadata with each document. It should retain as much as desired (and feasible) of the original functionality, look, and feel of each original document, while minimizing translation so as to minimize both labor and the potential for loss via corruption.

He also goes on to say that this ideal approach should offer alternatives of:

- safety
- quality
- volume of storage
- ease of access
- other attributes at varying costs
- importance of attributes for a given document, type of document, or body of documents
- single-step access (without layering)
- up front acceptance testing

Hedstrom (2000) argues that technological feasibility, cost-effectiveness, effectiveness, and acceptance are valid criteria for digital preservation strategies. These criteria also apply to access strategies. The primary concern of the Task Force on Archiving of Digital Information (TFADI) is to ensure "continued access indefinitely into the future of records stored in digital electronic form" (TFADI, 1996:iii).

Approaches to Maintaining Long-Term Access

Several methods—migration, refreshing, technology museums, etc—have been proposed by different researchers (TFADI, 1996; Willis, 1992; Rothenberg, 1998; MacCarn, 2000). Unfortunately, there is no agreed-upon single strategy that will satisfy all of the above criteria (TFADI, 1996; Kochtanek and Hein, 1999; Pace, 2000). With the

refinement of existing methods or development of new ones, a consensus may be reached on a viable solution. A survey of some of the methods that have been designed for long-term access is presented here.

Media-Based Approaches

Printed Hard Copy On Paper. Because paper tends to be more stable than magnetic, optical, and other electronic media (Lyman and Besser, 2000), some have proposed printing onto paper any of the information to be saved (TFADI, 1996). One of the major benefits of this strategy is that it requires little or no specialized hardware or software to retrieve the information depending on how it is printed. Another benefit of paper is that it can be made to archival quality specifications. This yields a much longer usable lifetime for the paper. A major flaw to printing everything out is that there is simply too much to print. "Printed documents of all kinds comprise only .003% of the total..." of information that is produced every year (Lyman and Varian, 2000:2). Another flaw is that printing the information explicitly puts it in a non-electronic medium that is time consuming to copy, manage, store, etc. The storage requirements for the roughly 250 megabytes worth of unique data for every man, woman, and child on the Earth (ibid.) would require 50 billion trees per year (Tennant, 1998).

Micrographics. This solution, promoted by Willis (1992) seeks to remedy the problem of requiring vast amounts of paper resources. This strategy is similar to printing information on paper except the medium is plastic and the information is miniaturized. Some of the major advantages are that it is already used as an archival medium with well-documented standards; it is easy to read the medium; and it can store a high resolution of

detail (ibid.). There are some disadvantages however. The medium has to be physically handled to access the information. It can become scratched in storage or use. The copy quality is degenerative (it loses about ten percent of resolution). There are also some problems with the transfer process (ibid.). Items "born digital" are those things that are created electronically and may not be directly transferable to microfilm. These digitally born objects may include video, audio, and databases as well as many other object types.

Nickel Slugs. Because paper and plastic tend to deteriorate when handled and are subject to limited environmental conditions, some have suggested engraving the digital information on nickel slugs (Rothenberg, 1998; Norsam Technologies, Inc., 2001). This, as well as any other method to transfer the information to non-electronic media, makes it far more difficult to access the information. Since the information is stored by engraving the information, these metal slugs offer the unique attribute of lasting for thousands of years. The storage capacity for the HD-ROM, produced by Norsam Technologies, Inc., is 200 gigabytes per disc and expandable to the petabyte range. At this rate though, it would take 10,000 of these discs just to preserve one year's worth of data. Not all strategies are medium dependent however.

Standards-Based Approaches

Several strategies regarding a standards-based solution have been posited. The major tenet of each of these strategies is that it is easier to maintain access to data if only one standard or a few standards are used. The Universal Preservation Format (UPF) was proposed by David MacCarn at the WGBH Education

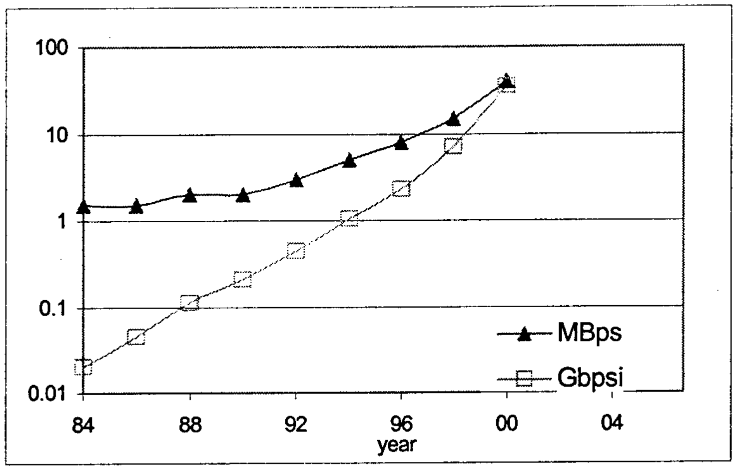
Foundation in Boston. The UPF is designed to reduce the confusion caused by the "veritable explosion of formats" (MacCarn, 2000:1). It also "specifies that machine-independent algorythyms be encapsulated within the stored media (MacCarn, 2000:1). Two strategies, the Bento Specification and the Open Media Format "are both media technologies that approach the UPF concept" (MacCarn, 2000:2). The major disadvantage of using a single format for storing all digital information is that "no computer technical standards have yet shown any likelihood of lasting forever--indeed most have become completely obsolete within a couple of software generations" (Bearman, 1999:3). The Time Capsule File System, proposed by Brian Zuzga (1995), is a similar approach to the Universal Preservation Format. It specifies a format that is "very similar to the RFC-822 format used for electronic mail" (Zuzga, 1995:16). It suffers from the same drawback as the UPF in that no single standard is likely to apply to technologies developed in twenty years and beyond. Some time into the future, scientists may find ways to process information based on an octal number system instead of binary.

Other Approaches

Technology Museums. Another approach is to store every generation of technology, keep the machines in working order, and run them with skilled operators. This is referred to as a technology museum. This approach would benefit by extending the longevity of computer systems and their original software to keep documents readable (Rothenberg, 1998). "Because originals are so important, [NARA has] a kind of museum

of equipment that will work or can be modified to work" (Carlin, 1998:2). This works on a small scale and for a short time, but for MIT which has a tremendous amount of information, the museum concept has not been as successful as MIT had hoped (Zuzga, 1995). A disadvantage of a technology museum is that "the hardware and software for digital media change so rapidly that it would be impossible to keep an up-to-date … museum" (Pace, 2000:56). Bearman (1999) agrees with Rothenberg (1998) that there are problems with technology museums. In fact, Rothenberg (1998:4) argues that "even if obsolete computers were stored carefully, maintained religiously, and never used, aging processes such as [metal migration and dopant diffusion] would eventually render them inoperative; using them routinely to access obsolete digital documents would undoubtedly accelerate their demise."

<u>Refreshing.</u> This strategy is the one that is probably most often employed (Pace, 2000). It "involves transferring digital materials to a new medium, for instance, changing from 5 ¼-inch floppies to CD-ROM, or from CD-ROM to DVD" (ibid.). NARA has procedures in place for refreshing. "Whenever any of the digital media in our custody shows signs of deterioration, or whenever they reach 10 years of age, we recopy the records to new media (Carlin, 1998:4). While this approach addresses the media instability problem, it does not fundamentally address formatting problems (Lyman and Besser, 2000). This method ensures that we will be able to access, for example, a Microsoft Word version 1 document, but without software to interpret the document, the file will be useless. The next strategy answers the interpretation problem.



The metrics are megabytes per second as the media spins (MBps) and gigabits per square inch (Gbpsi). (Adapted from Gray, *et al.*, 2000)

Figure 2: Magnetic Disk Parameters vs. Time

Refreshing is Inevitably Bound to Fail. "Disk capacity has improved 1,000 fold in the last 15 years, consistent with Moore's Law, but the transfer rate MBps has improved only 40x in the same time" (Gray, *et al.,* 2000:1). This means that disk capacity has a growth rate of 25:1 when compared with the transfer rate of data. The effect of this phenomenon is that our ability to store information is far exceeding our ability to transfer it to the next technological generation. Figure 2 shows how much faster the storage capacity has improved over the transfer rate increase.

Migration. Migration involves updating the format of the old digital object into what is currently used. Returning to the Word 1.0 document example, migration involves translating the information and storing it in Word 2000 format. This method is used frequently but

> the nearly universal experience has been that migration is labor-intensive, time-consuming, expensive, error-prone, and fraught with the danger of losing or corrupting information. Migration requires a unique new solution for each new format or paradigm and each type of document that is to be converted into that new form. Since every paradigm shift entails a new set of problems, there is not necessarily much to be learned from previous migration efforts, making each migration cycle just as difficult, expensive, and problematic as the last. Automatic conversion is rarely possible, and whether conversion is performed automatically, semiautomatically, or by hand, it is very likely to result in at least some loss or corruption, as documents are forced to fit into new forms (Rothenberg, 1998:6).

In other words, migrating all stored data to each new generation becomes increasingly infeasible, introduces the possibility for new losses (Rothenberg, 1998), and quickly borders on the impossible.

The Hybrid Approach. This strategy, proposed by Don Willis among others, suggests that for information that is not "born digital", preserving both the electronic version and a micrographic version mitigates the disadvantages of each individual method (Willis, 1992). This approach is not without its own, unique drawbacks. It would, in essence, triple the amount of already exponentially-growing information—one set of information would be the original, the second set the digital copy, and the third set the micrographic copy. Using standard compression methods, there are still about 240 terabytes of printed information yearly (Lyman and Varian, 2000). Even though this is a tremendous amount of information, it is a tiny amount of the total information produced yearly (ibid.).

Encapsulation. Rothenberg (1998) proposes that metadata and other information be encapsulated, or stored with, the digital information object. The other information would include the original executable software and operating system along with any other

pertinent data files. One factor that is both an advantage and a disadvantage is the inclusion of the software. On one hand, the encapsulated digital object would have the appropriate software to access the data. On the other hand, because every digital object would require its own copy of the software and operating system, it would require as many instances of the software as digital objects—even if all of the digital objects were at a single repository. Current operating systems require several hundred megabytes worth of hard drive space, and typical digital object software also requires hundreds of megabytes (Microsoft Corporation, 1999). Storing this nearly half of a gigabyte for one file that can range in the 10's of kilobytes seems inefficient. With the exponential data growth, maintaining individual copies of massive software sets seems infeasible.

Emulation. This strategy uses hardware and software emulators to access information stored on obsolete media and in obsolete formats. The information used to build both hardware and software emulators is very close to metaknowledge.

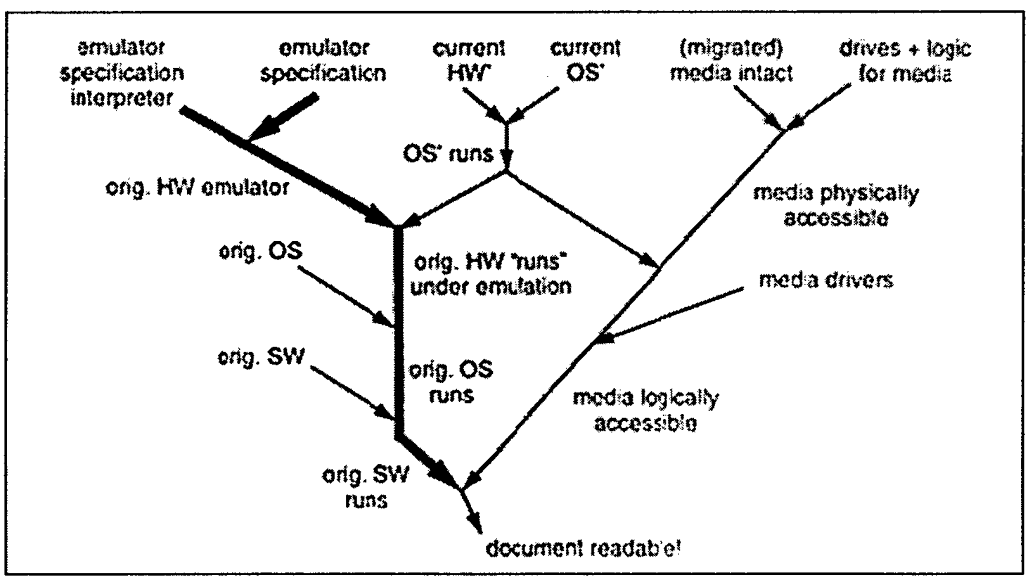


Figure 3: Emulation Chart (Rothenberg, 1998)

Rothenberg (1998) suggests that there are several benefits to the approach and argues that this is the best, if not the only, approach. The first benefit is that the emulators only need to be developed one time. This brings up the question, "What happens when the system the emulator runs on becomes obsolete?". A second benefit is that using the emulation of original software and hardware is the only way to accurately recreate the original digital environment. This will give the digital information object the same "look and feel" as it would have appeared using the obsolete technology. The negative aspect is that each emulation would also have to be maintained. Zuzga (1995:12) argues that there would be a "serious continuing cost" if emulation was used. One or more of these or some other strategies may well be put in place to help retain our digital heritage. However, even with this, there will be a need to recover stranded digital information. The DRS is designed to allow that type of recovery.

The Digital Rosetta Stone

Overview. In his thesis, Robertson (1996) explored the long-term access problem and suggested one approach to retrieving and interpreting data stored on obsolete media. Because Robertson's model was conceptual in nature, it did not include details of how best to implement it. This study will start with Robertson's model, and using the Delphi Technique to gather information from experts in the field, will explore the feasibility of this model, and add detail to its conceptual framework. The Digital Rosetta Stone Model was created by Robertson in 1996 as a way to maintain long-term access to static digital documents that were at risk of loss due to technological obsolescence.

Focus of the DRS. In keeping with the philosophy of the Digital Rosetta Stone, this study does not address accessing information in the short term. Recognizing that much stored information will be carried on to the next technological generation, the researcher does not focus on that area. Therefore, data migration problems of currently accessible devices are not under the DRS purview. The DRS does not attempt to recover information from media that has degraded beyond the point of data recognition either. The DRS was designed to be a last-ditch effort to recover stranded information. It is therefore to be used as a digital archaeology tool—recovering information that, until now, has been beyond reach. As such, the methods associated with a short-term perspective of maintaining access to information will not be discussed further. Digital archaeology—which is what the DRS is based on—is the bedrock by which all of the previous methods mentioned stake their potential (Pace, 2000).

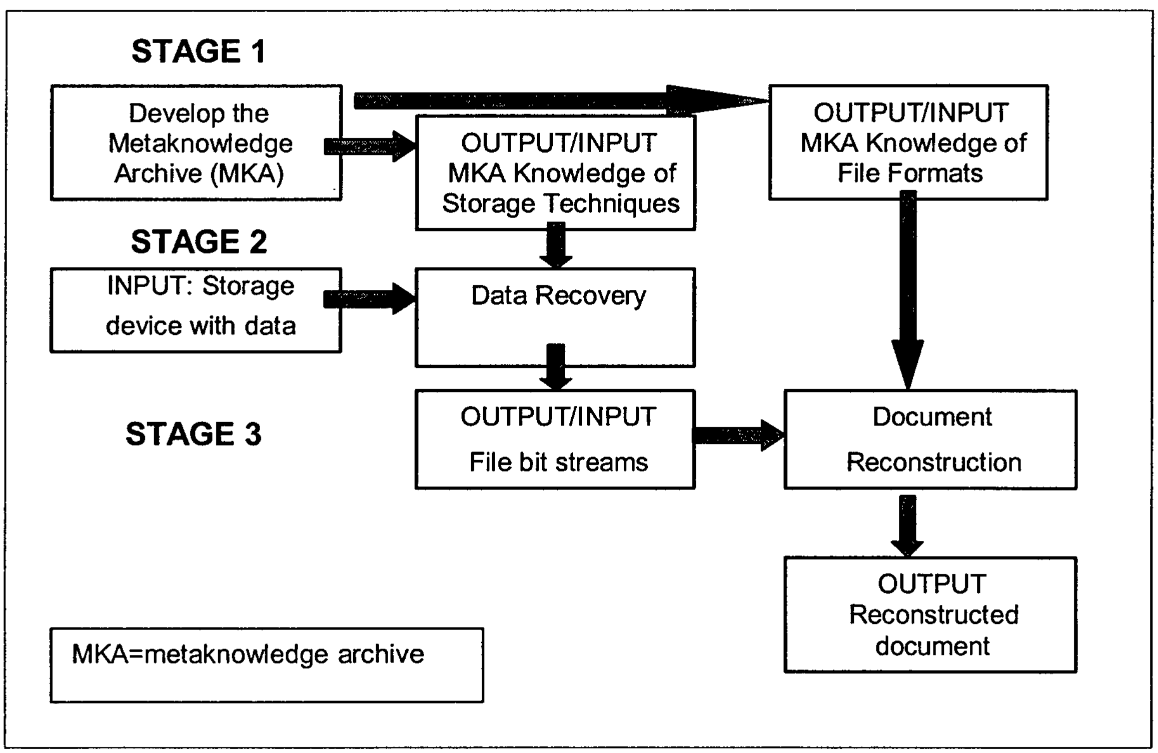


Figure 4: DRS Components (Robertson, 1996)

DRS Components

The DRS is composed of three major processes that are necessary to access digital information stored on obsolete storage devices or in obsolete formats—knowledge preservation, data recovery, and document reconstruction (Robertson, 1996). Developing each of these processes accurately is critical to the success of the DRS. The first major process, knowledge preservation, is addressed by the Metaknowledge Archive.

Metaknowledge Archive. Robertson (1996) proposed developing a repository of information necessary to both recover the data and reconstruct the document, which he calls a Metaknowledge Archive (MKA). This archive would be created through the act of knowledge preservation and would form the foundation for the other processes of the DRS Model. In fact, without this MKA, a file stored on an obsolete medium and/or in an obsolete format would be completely useless, even if the bits were preserved (Zuzga, 1995; Smith, 1998). Lyman and Besser (2000:14) point out that

> when we create or alter a digital object, we usually have much greater access to information about that object than at any other point in its life-cycle. Because we know so little about future viewing requirements, we don't know which of the seemingly innocuous bits of metadata [and metaknowledge] may later prove important to those environments. The more information we can save, the more likely we will be able to provide future generations with a "key" for unlocking the contents of whole classes of lost data.

Knowledge preservation is the process of collecting the information on the data storage and formatting techniques used by the designers and builders of information storage and processing devices. This includes the technical aspects of what constitutes a bit of information on this device, how it is arranged on the device, and how it is accessed. Information is also collected from systems and applications software that identifies the

file structures, along with all information necessary to recover and read the stored digital document. The MKA would be developed over time by a proposed Digital Rosetta Stone Office (DRSO), and would be made available to technicians to use to recover digital documents. The MKA obviates the need for the original designers and creators to be present for data recovery. The Defense Information Infrastructure Common Operating Environment (DII COE) developed and operated by the Defense Information Systems Agency (DISA) is one example of how software and hardware functionality is mapped out and categorized (Paige, 1997). Although it does not capture exactly the same information that would be in the MKA, the DII COE works much in the same fashion as the DRSO would in order to create and maintain the MKA. Rothenberg (1998), Bearman (1999), and Pace (2000) all stress that the effort necessary for an effective MKA is significant. Storage techniques can be quite complex, not to mention that current and cutting-edge technologies are fiercely protected—companies stake out and defend proprietary market advantages to protect their profit (Rivette, *et al.*, 2000). One side benefit of developing the MKA is that it could also help in other cases by providing information on standards or hardware and software used and pointing people to places where appropriate hardware and software can be found.

Reproducing the Bitstream. Armed with the knowledge of storage techniques, recovery technicians can begin. Data recovery is the process of retrieving the bitstream from the outdated and obsolete medium and moving it to a current storage device. If necessary, the information in the MKA could be used to create a new medium access device. The access method may be altogether different than the original device used. For instance, instead of building a CD-ROM drive to recover a bitstream, the DRSO workers

might use a high-resolution scan of the CD and software to interpret the image (Robertson, 1996). This may help if the media is fragile and may not survive traditional data access methods.

Interpreting the Bitstream. Once the bitstream is accessible to the modern computing environment, document reconstruction can take place. This is where the bitstream—manipulated using the knowledge of formatting techniques—is displayed as the original digital information object. Depending on how well the MKA has accurately and thoroughly captured all of the formatting techniques, the reconstructed document can be an exact representation of the original document.

Output. The result of going through each of the stages of the DRS would result in a recovered digital information object. Given the variety of file formats, the reconstructed object could be an encapsulated document containing metadata or a simple ASCII-text file. This flexibility gives the DRS the hardiness to be a long-term solution.

Benefits. Such a "universal solution to a ubiquitous problem could consolidate the market for capture, storage, and maintenance technologies" (Hedstrom, 2000:1). If fully developed, the DRS could prevent untold losses of information. The TFADI has suggested that a fail-safe mechanism be created (TFADI, 1996) and the DRS has the potential to satisfy this criterion.

The DRS picks up where the preservation strategies leave off. It is designed to handle any number of formats, either existing or future ones. This would allow people to understand what the old format consisted of and determine the best way to interpret the bitstream based on the current computing environment. If the digital artifact was an

encapsulated digital object, the encapsulated information could be read as well as the information.

Summary

Because most of the strategies discussed here are primarily focused on preservation, they do not address the problem to be dealt with by the DRS. Digital archaeology, by its very nature assumes that preservation of information has occurred. If the data no longer exist, then nothing will be able to bring them back. The DRS is uniquely different because it focuses on retrieving a bitstream from an obsolete medium and interpreting that bitstream so the original information can be displayed. It does this by using the MKA. The next chapter describes the manner by which the rest of the research was conducted.

III. Methodology

Introduction

The methodology chapter describes how the research for this thesis was structured and performed. This research is inductive and qualitative in nature. That is, it seeks to develop an understanding of a topic rather than test a theory. The qualitative side deals with opinion statements leading to generalizations. Thus, the methods used do not rely heavily on statistics, although some statistic measures were taken.

Overview of the Delphi Technique

The methodology for this research was a Delphi Study. The Delphi Technique was developed in the 1950s by Olaf Helmer and Norman Dalkey, scientists at the RAND Corporation (Linstone & Turoff, 1975). It was initially used as a long-range forecasting tool but has since developed to include a number of other uses. It involves a group of experts who provide their opinion on a certain topic. The ideas generated are then analyzed and condensed to determine a level of consensus. The Delphi Technique is performed in a series of rounds with experts. It solicits ideas and fosters discussion about them. The experts then provide opinions about the statements. These opinion are analyzed to determine if a group consensus exists. This iterative process of rounds and analysis continues until a consensus or stabilization point has been reached. Stabilization indicates that inter-round answers have not changed beyond an appreciable amount. The opinions are annotated using a Likert-type scale ranging from Strongly Disagree to Strongly Agree and for Very Important to Not Important (Linstone & Turoff, 1975;

Kochtanek and Hein, 1999). Because this research is inductive in nature, group consensus will not be the only measure of "success". The idea generation in and of itself will also be useful to the DRS—the ideas submitted by experts in the field can provide important insights into the strengths and weaknesses of the model.

Advantages	Disadvantages
More information and knowledge are available	The process takes longer than individual decision making, so it is costlier
More alternatives are likely to be generated	Compromise decisions resulting from indecisiveness may occur
More acceptance of the final decision is likely	One person may dominate the group
Enhanced communication of the decision may result	Groupthink may occur
Better decisions generally emerge	

Table 2: Advantages and Disadvantages of the Delphi Technique (Griffin, 1999:281)

Having briefly explained the Delphi Technique, the next section will describe this particular implementation of it. Utilizing electronic mail made using the Delphi Technique far more practical than postal mail because of the time constraints. The nature of the data was primarily qualitative due to this being a grounded theory study.

The DAE Population

The population of interest in this study was the group of people I call digital archivist experts (DAEs), whose knowledge about the subject area is key to exploring the potential of the DRS Model. Those who constitute the digital archivist community include Information Technology (IT) specialists who are responsible for maintaining long-term access to digital information and are primarily librarians, digital archivists, and

academicians. Individuals in this community may be found in a wide variety of industries and government agencies. Key technology makers were considered because of their impact on technology. The organizations under consideration were asked to decide who was most suited to participate in this study.

The Participants

During the literature review, potential participants were identified based on the types of articles they wrote or other demonstrated capacity. The organizations that were contacted are noted in Table 3.

1. The University of Pittsburgh, School of Information Science	2. The Syracuse University Library
3. Bellcore	4. WGBH
5. The United States Air Force Historical Research Agency	6. The RAND Corporation
7. Connectex	8. The National Archives and Records Administration
9. The United Kingdom Office for Library and Information Networking	10. Microsoft
11. INSO	12. SUN Microsystems
13. The Library of Congress	14. United States Army
15. United States Navy	16. Defense Technical Information Center
17. The University of Michigan	18. The Preservation Services Group at The Research Library Group
19. Yale University, Preservation Department	20. IOMEGA Corporation

Table 3: Organizations Contacted for Participation

The organizations that participated in this study are listed from 1 through 8. Some experts worked in groups in their respective organizations to develop the answers. The result is that as many as 12 to 15 people actually contributed to this study. For purposes

of the technique, the individuals in the group knew who was participating but did not know who made what comments – anonymity has been shown to increase creativity and idea generation (Linstone and Turoff, 1975).

Non-Probability Sampling. The participants were selected based on their perceived qualifications. Because this study is not attempting to use statistics to make inferences about a larger population, non-probability sampling can be used (Dooley, 1995).

Request for Participation

Letters about the study were sent to the prospective participants. A sample of these letters is included as Appendix A. A thank-you letter was sent to those who agreed to participate. It is also included as Appendix B. Along with that letter, a number of other documents were attached. Included in the initial package, Appendix B attachments, were a purpose of intent statement including all the information about who is performing the research and how to contact them, an executive summary of the problem, a detailed problem statement, a simplified version of the Digital Rosetta Stone Model, as well as Heminger and Robertson's paper, as it was published in the Communications of the AIS; and a description of the Delphi Technique.

Preparing for the First Round

The literature review provided a glimpse into the advantages and disadvantages of different preservation strategies. Based on this, the researcher developed several research

questions. They were designed to elicit responses regarding the nature of the DRS and how well it could work. These questions addressed the strengths and weaknesses of the model, as well as how it fit in to the overall preservation and access environment.

Pilot Project

The goal of this portion of the research was to refine the questions and determine whether the material included in the initial package was sufficient and satisfactory. Several graduate students at AFIT were given a chance to participate in the pilot project. Their comments about the questions, attachments, and answers to the questions helped to determine what to include and how to word the questions. An informational package about the DRS was developed because the experts were not expected to know the details of the concept.

The First Round

For the first round, a description of the DRS Model was sent out to the group, along with instructions on how to participate and a fuller explanation of the study's intent. The question topic asked of the group pertained to the DRS Model and its appropriateness, as well as, its completeness for maintaining long-term access to digital objects. The Request for Participation packet and each rounds' packets are included as appendices. The goal of this round was to generate as many ideas about the DRS as the experts felt appropriate. These ideas formed the basis for beginning to develop a consensus.

Respondent's Response Time

It was expected that the response time would be one and a half weeks and was identified as the time limit for the members. It was later expanded to three weeks when it became apparent that the busier members needed more time. This was designed to allow a reasonable amount time for busy members to finish with their input and reply. For those that did not respond by the end of the time period, an attempt was made to contact them via either telephone or email to see if there was a problem. The members were still included in the study, even if they did not respond. This was done primarily as a means to minimize attrition. The actual time period of five weeks was much greater than anticipated.

The Second Round

The second round documentation, included as Appendix E, included a review of the first round answers, clarifications to the model, and ideas in tabular form for the experts to comment on. The members of the Delphi group were asked to agree or disagree, on a five-point Likert Scale with the condensed results and to refine their statements regarding the DRS Model.

1	2	3	4	5
Not Important	←	**Importance**	→	Very Important

Figure 5: Importance Scale

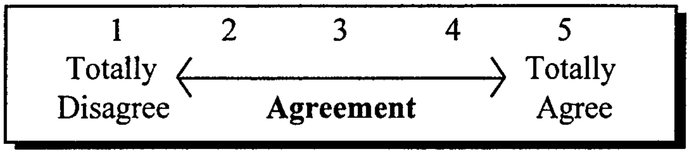


Figure 6: Agreement Scale

<u>Consensus.</u> Consensus is a measure of how much people agree with one another. For this Delphi Assessment, consensus for agreement and high importance was defined as a median of one-half a point above the middle of the Likert scale or higher. Consensus for disagreement or low importance was defined as a median of one-half a point below the middle of the Likert scale or lower. A group consensus that was in the middle of the Likert scale resulted in an unsure rating, for either importance or agreement. There was no distinction between a group consensus of unsure and non-consensus. Non-consensus was for all intents and purposes defined as a consensus of unsure. The median is a useful measure of central tendency—the best representation of a group of responses—because it "reflects the middle value" of responses and it "takes into account all of the observations" (Dooley, 1995:21).

Second Round Response Time and Analysis

The response time for the second round was to be two weeks with analysis to begin with the first response and end soon after the last response. All of the members did not respond in the expected time. After another week, a follow-up email was sent to identify problems or questions. To encourage higher participation, several AFIT students, who had participated in the pilot project and were familiar with the topic, were asked to

participate in a time trial. Based on the time it took for them to complete the second round, a follow-up call was made after another week and the participants were instructed that it should take somewhere around 30 minutes to complete. They indicated that they would try to work on it and send their answers in soon.

The Third Round

This round was expected to be the final round due to research time constraints. This round consisted of sending out the Second Round Report, included as Appendix G. The group was asked to comment on it for accuracy and completeness and to state their overall assessment of it. The response time was set at one week. The majority of responses were received in that time frame. One more response was received on the eighth day. The results from this round formed the basis of answers to the research questions. The researcher drew some conclusions in Chapter 5 regarding the next step for the DRS Model.

Summary

This research seeks to develop a body of expert opinion and possibly develop a consensus on the DRS Model and determine what the next step for it should be. A knowledgeable group of people familiar with the access and preservation environment was found during the literature review and was selected for participation. Sending out an information package to familiarize the experts facilitated the iterative rounds of discussion based on the Delphi Technique. Based on an analysis of all the data gathered

in the rounds, implications of the research were developed. This analysis also helped determine what to do with the DRS Model. The next chapter, Chapter 4, deals with the results of each round and a corresponding analysis.

IV. Results and Analysis

Introduction

The Delphi Technique was utilized for this research. There were three rounds of discussion and feedback. The following material was developed by each of the three rounds. An analysis of each round was performed and was used as the input for the next round. Round 3 was the last round and the results from it led to the development of answers for the research questions.

First Round Results and Analysis

The results from the first round, in Appendix D, were analyzed using content analysis. Content analysis is the procedure for measuring the occurrences of selected lexical or vocabulary features in speech or text (Dooley, 1995). In other words, content analysis means looking at different statements based on the intent and determining if they were similar in nature. A major advantage of content analysis is that if done well, it should be replicable (Krippendorff, 1980). The data is included for further reference and verification. The resulting generalizations from the first round statements formed the basis of discussion for the rest of the study.

Round One Report. Some of the following are examples of statements made by the experts in the first round. In regards to the strengths of the DRS Model "[It] does a good job of describing the characteristics and attributes of electronic files that affect preservation and access. It lays out a methodology to maintain the ability to reliably retrieve and reconstruct digital documents" (Expert B). "Unlike most approaches, it also

has the potential for allowing obsolete digital storage media to be read in the future, even if no readers for such media still exist" (Expert D).

The second question asked the experts to comment on the areas of the model that need improvement. "I believe the repository of metadata about the world's data is needed, but the data itself should be distributed, (Depository Libraries). Finally the mechanics of reading should consist of a device that is unlikely to change drastically, or become obsolete.......the human eye" (Expert A). "The model adequately addresses digital files that already exist but needs to provide a workable solution for the future – a standard format for document creation and markup" (Expert B). All of the comments, including these just listed, can be found in Appendix D. One statement was misunderstood by the researcher and instead of being stated as the DRS needed to address "self-describing media", it was reported to the group as a need for "self-describing metadata". Therefore the original statement was not commented on by the group during the second round.

Appendix E is the next document that was sent to the group of experts. It was created in response to the first round statements. It is a report of the first round and contains the clarifications to experts' misconceptions as well as the second round topics.

Summary of Round One Responses. The response rate for the for the first round is as follows:

No response received.........................2 (22.2%)
Responses received...........................7 (77.8%)
Number of total statements received........66
Number of unique statements received.....54

Second Round Results and Analysis

The opinions expressed during the second round were predominantly recorded in the form of Likert-type scales, as shown in Appendix E, Section III. At the end of each question, however, there was room for additional comments. The opinions from the different experts are listed in Appendix F.

Round Two Report. These findings are what I gathered from the responses from the experts. I categorized the statements into eight areas or topics that each statement seemed to address. They are ordered in a manner that tries to present an overall picture of the DRS landscape. A matrix of categories for opinions was also developed. This facilitated categorization of each of the statements based on the level of consensus on statement importance and statement agreement. Before the Round 2 report was sent, one participant, who had not responded at all, decided to end involvement with the study because of his workload.

Statement Topics. The first topic deals with the preservation and access environment that created the need for the DRS. The second topic deals with physical media devices and digital objects. The third topic covers relevant areas of the development of the Digital Rosetta Stone. The fourth covers the focus of the DRS. The fifth topic covers the methodology of the DRS and the following two areas, six and seven, go into more detail of the methodology category. The eighth, and last, area deals with statements made about the DRS implementation details. These topics are designed to give the reader some idea about where each of the statements belong in the DRS landscape.

Statement Topics

1. Preservation and Access Environment
2. Media and Digital Objects
3. Development of the DRS
4. DRS Focus
5. DRS Methodology
6. Metaknowledge Archive
7. Software, Logical Formats and Physical Formats
8. DRS Implementation Details

The experts submitted opinions about the statements in the form of two parts. The first opinion was directly related to whether or not the expert agreed with the statement. The second opinion dealt with whether or not the statement was important to the DRS. The opinions were recorded using a 5-point Likert-type scale, with the low end being either disagree or not important. High numbers were used to indicate agreement or high importance. Question 5 related to assumptions that the DRS made. The experts were also asked to state if these assumptions regarding Question 5 were valid or not. Each of the statements has two opinion parts: statement agreement and statement importance. Each of the opinion parts has three possible answers: Agree/Important, Unsure/Unsure, or Disagree/Unimportant. This results in nine possible statement agreement and statement importance opinion outcomes or categories.

		Levels of Importance		
		High (A)	Unsure (B)	Low (C)
Levels of Agreement	High (1)	Important and Agree	Unsure Important and Agree	Not Important and Agree
	Unsure (2)	Important and Unsure Agree	Unsure Important and Unsure Agree	Not Important and Unsure Agree
	Low (3)	Important and Disagree	Unsure Important and Disagree	Not Important and Disagree

Table 4: Categories for Opinions

For purposes of tracking which statements belong in what category, each row and column has been labeled with a letter or number, in addition to the level of importance or agreement. The Importance Level columns have been labeled A, B, and C, corresponding to their order. The Agreement Level rows have been labeled with 1, 2, and 3. For example, the category of Important and Agree will be referenced as Category A1. The Important and Disagree category will be referred to as Category A3. Also, each one of the eight statement topics will be referred to by its corresponding number. Every opinion discussed in this report will have a similar heading consisting of the category rating (A1, A2, A3, B1, etc.) and statement topic number (1-8). In the case of the first opinion, the heading will be "A1.1 Preservation and Access Environment"—A1 being the category for the Important and Agree opinions.

Not every one of the nine categories for opinions had every statement topic in it, but all of the topics fit into the categories. The statement topics will be discussed by level of importance followed by level of agreement.

Group Rating of Unsure Versus Disagree or Not Important. There is a fine distinction that needs to be made between a rating of Unsure and a rating of Disagree or Not Important. For instance, the group could come to a consensus on a statement--deciding that it was important but disagree with it. This disagreement should not be confused with not having a consensus. If all of the experts said they disagreed with a statement, then the group would have come to a consensus that they, as a whole, disagreed with a statement. The points where the group did not come to a consensus,

either for importance or agreement, are listed as Unsure. Also, a statement could be listed in the Unsure category based on a group consensus of unsure.

Discussion of the Group's Opinions on Each of the Statements

A1. Important and Agree Category

This category consists of those topics on which the group of experts reached a consensus that they agree with the statements and also agree that the statements were materially important to the Digital Rosetta Stone and its development. One third of the statements fell in this category. The statement that had been mis-reported in Round Two regarding the need for "self-describing metadata" was corrected in this report to read "self-describing media".

A1.1 Preservation and Access Environment. As young as the digital world is, we are already seeing that there is a definite need for digital archaeology. This validates the DRS assumption of a need for digital archaeology. If the DRS is to be successful, it needs to be aware of other strategies for long-term access and those for preservation as well as be compatible with them.

A1.2 Media and Digital Objects. Making sure that the output matches the original is important. The developers of the DRS need to take this into account. Because of the long-term nature of the DRS and the general instability of media, the DRS should seek to use or develop methods to handle media degradation and failure. To aid in future recovery efforts, the developers should address the need for self-describing media, although the DRS does not currently do this. To the extent that this could be done,

utilizing self-describing media would certainly simplify the DRS. It would assist in the process of recovering the bitstream, leaving only the interpretation of the bitstream to complete document recovery.

A1.3-5. There were no statements in the A1.3-5 categories.

A1.6 Metaknowledge Archive. The DRS can accomplish its long-term access mission because it maintains the Metaknowledge Archive. Because the foundation of the DRS is the MKA, the criteria for the MKA needs to be developed further and clearly specified. This statement has an important caveat. It has not been shown that an MKA, populated with the required information can be built because there is no "accepted or demonstrated methodology for creating that required metaknowledge, and there is much evidence to indicate that this may be far more difficult than it sounds" (Expert F, Round 3 Comments).

A1.7 Software, Logical Formats and Physical Formats. Software is very important, and a concerted effort with software developers will be necessary to capture sufficient information to assist in recovery efforts. Some files are application independent, such as .jpeg or .bmp. The "native format" is the format that the originating software used for the file and this format is important to understand. Some of these digital documents will be textual or paper like, but the rest will not. Because the example used was of a text-only document, the group wanted to make sure that the model would attempt to recover the non-textual digital objects was clear. Examples of these non-text digital information object types are database files, graphics, and encapsulated metadata digital objects.

A1.8 DRS Implementation Details. The group strongly agrees that maintaining long-term access to documents is important and that the DRS allows for that access even if no readers exist for that medium. The sentiment was not unanimous—there was one who disagreed on the DRS portion of the statement. Cooperation for implementing the DRS with the public and private sectors is necessary. The development process should include a prototype to determine technical feasibility, total life-cycle cost analysis, and a probability determination of a successful DRS implementation. A consortium of those who store and use information needs to be developed to further build the model. To help get the process of DRS development going, it needs to be exposed to others where substantive work is being done in this field.

A2. Important but Unsure of Agreement Category

These issues are important to the DRS but the experts are not sure if they agree with the items or not.

A2.1 Preservation and Access Environment. Addressing the fundamental issues of technically translating documents over time is important, but the experts are unsure that the DRS does this. At this point in its infancy, the DRS does not yet actually cover the technical issues. This can be addressed as the model is developed.

A2.2 Media and Digital Objects. Media instability is an important problem, but the group is unsure if the DRS addresses that problem. Data about the original storage media are important, but the group is unsure that the data will be available when it comes time to capture it for the MKA. The group is not sure that the DRS makes the assumption

that this data will be available. One expert says that it is easier to capture the data when it is readily available.

A2.3-6. There were no statements in the A2.3-6 categories.

A2.7 Software, Logical Formats and Physical Formats. It is the format of the logical bitstream that is important to the software and how the data is presented—not the actual storage mechanism of the bits, but the group is unsure if the DRS fully addresses this.

The group is not sure if the DRS makes the assumption that the physical format of the digital artifact's logical bitstream is more important than the logical bitstream—the bitstream after it has been retrieved from the storage device. They do not think that the physical format is more important than the logical bitstream. In other words, both the physical format and the software formats are important to data recovery.

A2.8. There were no statements in the A2.8 category.

A3. Important and Disagree Category

This grouping of items was deemed to be important to the DRS by the Delphi group, but they disagreed with the statements. This suggests a consistency in responses, because some of the statements were relatively opposite with what some of the agree statements were.

A3.1-4. There were no statements in the A3.1-4 categories.

A3.5 DRS Methodology. The group thinks a methodology to maintain the ability to reliably retrieve and reconstruct digital documents is important but they do not think

that the DRS has such a methodology. It could be that they do not think it has or that it will not have one at all. The intention of the DRS is to develop the methodology, but it is not currently in place. They agree that adequate resources are necessary but do not think that the DRS assumes that the needed resources will be available. The group also came to the conclusion that the DRS is important and does warrant significant investigation at this time.

A3.6 Metaknowledge Archive. The group agrees that the MKA is important but is unsure if the MKA will be available. They do not think that the DRS makes this assumption. Preserved documents are important but do not necessarily meet preservation criteria. The group does not think the DRS makes this assumption either. Media metaknowledge standards are important, but are not adhered to or valid. The group does not think the DRS makes this assumption.

A3.7 Software, Logical Formats and Physical Formats. The group thinks that software behavior and physical format are important but that the DRS should not focus more on the software behavior than the physical format. Data re-creation is important but knowledge preservation, data recovery, and document reconstruction are not all that is needed. They also do not think that the DRS makes this assumption. A digital document's meaning is important but not entirely conveyed by the bitstream. They agree that the DRS does not make this assumption.

A3.8. There were no statements in the A3.8 category.

B1. Unsure Important and Agree Category

The experts were unsure of how important these items were to the DRS but did reach a consensus on agreement for each item.

B1.1 Preservation and Access Environment. The DRS is in agreement with Rothenberg's emulation-based strategy in that it recognizes the importance of retaining original formats. They also agree that it diverges in the fact that the emulators are used in Rothenberg's solution to properly interpret the bitstream, but not in the DRS. They are not sure how important this statement is to the DRS.

It differs from Persistent Object Preservation because the DRS is an access method not a preservation method. Because it does differ, the group is unclear on how important Persistent Object Preservation is in terms of impact on the DRS.

B1.2. There were no statements in the B1.2 category.

B1.3 Development of the DRS. They agree that the government should help undertake the implementation of the DRS but are not sure how important or to what level the government should have its involvement.

B1.4 DRS Focus. The group agrees that the DRS recognizes the importance of the digital object's original characteristics, but rates the importance as "unsure".

B1.5 DRS Methodology. The DRS needs to spell out a methodology for commercial cooperation, but the group is unsure how important it is to the overall success of the DRS. They agree that it needs to have an analysis of the cost-effectiveness of other approaches. This goes to the overall awareness of the other methods as stated previously.

B1.6 Metaknowledge Archive. The metaknowledge should be accumulated, however, the group is unsure how this will affect the overall implementation of the DRS.

B1.7-8. There were no statements in the B1.7-8 categories.

B2. Unsure Important and Unsure Agree Category

The group was unsure of how important these items are to the DRS and are ambivalent about whether or not the group agrees with these statements.

B2.1 Preservation and Access Environment. The group was unsure of how the DRS compared to a hybrid systems approach for preservation of printed materials and was also not sure how this applied to the DRS. The group was unsure of whether the DRS was similar to the Universal Preservation Format. This may suggest that not all of the experts were familiar with the UPF.

B2.2-4. There were no statements in the B2 category.

B2.5 DRS Methodology. The group was unsure of whether the MKA should be distributed or centralized. They were also unsure of how important the level of centralization or decentralization was to the DRS. They were unsure of whether it needed to develop functional standards for chronological interoperability. They were also unsure of how important this was to the DRS. This might be explained as the experts not being clear on the exact meaning of "functional standards for chronological interoperability".

B2.6-8. There were no statements in the B2.6-8 categories.

B3. Unsure Important and Disagree Category

The group is unsure of how these items relate to the DRS but disagree with the statements as a whole.

B3.1-3. There were no statements in the B3.1-3 categories.

B3.4 DRS Focus. The DRS does not assume that digital archiving is solely a technological problem. The experts are unsure of how important this is.

B3.5. There were no statements in the B3.5 category.

B3.6 Metaknowledge Archive. Media metaknowledge is not rigidly defined before coming to market but the group does not see how this applies to the DRS. They do not think the DRS makes this assumption.

B3.7-8. There were no statements in the B3.7-8 categories.

C1. Not Important and Agree Category

The group did not think these items directly affected the DRS but did agree on them.

C1.1. There were no statements in the C1.1 category.

C1.2 Media and Digital Objects. The DRS does not address what to do with the data after recovery. This is not important, as one expert stated "The DRS is concerned with data recovery not what happens to the data after recovery." In other words, let the people who wanted the data in the first place decide what they will do with it. The DRS does not address the context or order of a document in a collection and this fact is not important.

C1.3-8. There were no statements in the C1.3-8 categories.

C2. Not Important and Unsure Agree Category

These items are not important and the experts cannot be sure if they agree with the statements.

C2.1-3. There were no statements in the C2.1-3 categories.

C2.4 DRS Focus. The DRS may lack the archival distinction between a document and a record, but it does not really matter. The DRS may not address legal-related issues such as intellectual property and is not important that it does not do this. The group seems to be evenly split on the importance level of this statement. The statement might have some applicability if further clarified.

C2.5-8. There were no statements in the C2.5-8 categories.

C3. Not Important and Disagree Category

These items are not important to the DRS and the group disagrees with the statements.

C3.1-3. There were no statements in the C3.1-3 categories.

C3.4 DRS Focus. The DRS does not have too narrow a view of what constitutes data recovery, but this is not too important.

C3.5-8. There were no statements in the C3.5-8 categories.

This marks the end of the second round report.

Summary of Second Round Responses. The response rate for the for the second round is as follows:

No response received.......................1 (11.1%)
Responses received.........................8 (88.9%)
Number of Importance Opinions..........385
Number of Agreement Opinions..........374
Number of Validity Opinions.............102

Third Round Results and Analysis

The third round consisted of sending the Second Round Report to the group. The group was requested to review the report and comment on any portion of the report that it felt was appropriate. They were asked to see if the generalizations made sense and were reasonable assessments of the second round opinions. Each of the expert's opinions are listed in Appendix H. Overall, the group responded positively to the Second Round Report. There were a few minor questions and some statements made regarding the need for clarity on some of the categories. Also, one respondent stressed that for the thesis, certain terms, such as digital archaeology and access, needed to be well defined so as to not confuse or mislead readers. Digital archaeology is "an approach that relies almost totally on future efforts to decipher saved digital bitstreams" (Expert F, Round 3 Comments).

The round three responses indicated a high approval of the round two report, which validates the use of a median discriminator value of half of a point above or below the middle of the Likert scale. If the group had not come to such an overall agreement,

the median value used to distinguish levels of agreement and importance could have been called into question.

Summary of Third Round Responses. The response rate for the for the third round is as follows:

No response received........................2 (22.2%)
Responses received...........................7 (77.8%)

Overall Response Rates

There were three participants who had originally agreed to participate but did not take part in any round. They were not included in the response rates. Of all of the participants, four participated in every round. Five took part in two rounds. No one participated in just one round. Overall, nine experts participated in this study at one point or another.

Research Questions Answered

The purpose of this research is to answer the research questions and based on the answers, develop recommendations for the future of the DRS. The following is the discussion of the research questions' answers and Chapter 5 contains the recommendations. The answers are derived from all three rounds. The statements that were listed in the "Unsure Importance" or "Unimportant" categories are not listed. The assumptions that the DRS does not make are found in the "Disagree" category and are not listed here.

Research Question 1: What are the strengths of the Digital Rosetta Stone Model?

- It recognizes the importance of retaining access to objects even as the technology for storing them becomes obsolete.
- It allows for access even if no readers for such a medium exists.
- It has the idea of a central registration of document types and specifications.

Research Question 2: What are the areas in the Digital Rosetta Stone Model that need improvement?

- Where possible, the DRS should integrate well with archiving.
- It does not describe how to handle media degradation and media failure.
- The Metaknowledge criteria needs to be further developed.
- The DRS should place an equal emphasis on the behavior of the software during interpretation of the bitstream and the retrieval process from the physical medium.

Research Question 3: What is missing from the Digital Rosetta Stone Model?

- The awareness of other long-term access efforts and its compatibility with them.
- The need for self-describing media.
- It does not address the problem of authenticity, or integrity, of the original document.
- It does not address verification and validation of the translation.
- It misses the importance that software plays in interpreting the digital documents by the fact that the behavior of such software is not implicit in a digital artifact's format.

Research Question 4: How does the Digital Rosetta Stone Model compare with other models in relation to maintaining long-term access to digital documents?

- Other schemas are geared toward digital document preservation.

Research Question 5: What are the underlying assumptions of the Digital Rosetta Stone Model? If the DRS makes the assumption, is it valid? All assumptions made by the DRS were valid.

- The DRS assumes we are in a situation that needs digital archaeology.
- The "native format" is what the original application created.
- Some preserved digital documents will be textual.
- Cooperation with the public and private sectors is necessary.

Research Question 6: What steps are necessary to begin implementation of the Digital Rosetta Stone Model?

- Clarify whether the model depends on the original medium being available at the time of need.
- Assuming we are ready for a decision, clarify how the model would attempt to recover non-textual information.
- Assuming a feasibility study has been performed, consider the total life cycle costs and probability of the model being successfully implemented.
- Development of the consortium to further build the model.
- The DRS warrants significant investigation at this time.

Research Question 7: Who should undertake development and implementation of the Digital Rosetta Stone? And why?

- A consortium of those who use and store information.

Research Question 8: Do the experts have anything else to contribute that does not fit in the previous questions?

- This project needs to be brought into the contact of others where substantive work in this field is being done.

Summary

This chapter has covered each of the rounds and their resulting analyses. Based on each analysis, the answers to the research questions were established. The statements

made in round one that were later determined to be of questionable importance or low importance were not listed as answers. Even those statements that were agreed upon, but found to not be relevant, were not listed. While those statements may be interesting in and of themselves, the group did not find them directly applicable to the DRS. As stated earlier, Chapter 5 covers the implications of this research.

V. Discussion and Recommendations

Chapter Overview

This chapter discusses the findings and presents the conclusion of this thesis research. It discusses the problem of maintaining long-term access to static digital documents and related implications to the preservation and access community including the United States Air Force. There are some limitations to this study and they are also addressed. Based on this research, recommendations for future researchers are made.

Discussion

This thesis represents the first assessment of the Digital Rosetta Stone Model by the expert community. As recommended by Robertson (1996), it presented the model to the archival community and other interested parties. Some of their overall statements in the first round suggested that the overall impression of the DRS was negative. They addressed problems with the practicality of such an undertaking and that the DRS's focus may be misguided. However, when asked to address the research questions, their answers proved to be realistic but hopeful.

<u>Research Question 1: What are the strengths of the Digital Rosetta Stone Model?</u> The DRS is designed to be the link between viewable information and data stored in obsolete hardware and/or software technologies. The DRS implements two steps necessary to retrieve the information. When one tries to recover information, it does not matter how long the technology has been obsolete, from a technological viewpoint, because the MKA, the repository of technological information, should have everything

necessary to recover the information. Typically, the longer a technology has not been used, the harder it is to try to understand the intimate details of its inner working. The DRS is intended to reduce the effect that time has on understanding such details.

Research Question 2: What are the areas in the Digital Rosetta Stone Model that need improvement? The preservation and access community should continue the concepts of the DRS with existing and potential preservation strategies in mind. Symbiosis between these preservation strategies and the DRS could then be nurtured. If the idea of a DRS is accepted by the creators and maintainers of hardware and software technologies, they could engage in populating the MKA with metaknowledge. The success of the DRS is entirely dependent on the right information being in the MKA. If the MKA does not contain everything necessary to retrieve the bitstream and then interpret it to display the stored information, the recovery process will become difficult, if not impossible.

If the stored bits of information do not survive, then the DRS is useless. Knowing the different environmental storage requirements of the different media could help DRS technicians know how best to handle and store the media until bitstream retrieval has occurred. Oftentimes, ignorance in the handling of sensitive objects can undo years of preservation. This is one area where working with preservation specialists could reduce the amount of media failure and slow the process of media degradation caused by abuse or neglect.

Research Question 3: What is missing from the Digital Rosetta Stone Model? In developing the DRS, the experiences gained from others who have worked on access

strategies needs to be taken into account. This could help overcome unforeseen obstacles in DRS development or help to build a more robust DRS.

Self-describing media could help fill in for the MKA if there are any gaps or inconsistencies and could reduce reliance on a concept such as the DRS. This is where the medium itself has written instructions on how to recover the information. As one expert commented on the example regarding the 8-track punched paper tape, the encoding information could have been written on the other side. This becomes more difficult when, as for example, a Digital Video Disc (DVD) has practically no room for displaying any human-readable information—both sides are used for data storage, as compared to a regular compact disc that uses one side for written information.

When reconstructing the original document, unless there is a human-readable copy or other known stored instance of it, verifying that the output is exactly the same as when viewed using the "native software" will be difficult. One of the best ways to ensure that the DRS produces the correct output is to test it on stored digital information that is not yet obsolete.

The DRS developers need to work with software designers to identify what software behavior is not contained in the digital object's format. This, as yet unacknowledged behavior, could change the way the bitstream is interpreted and not produce the intended result. The DRS should not attempt to re-create the full functionality of the native software, only enough that is necessary to properly display the stored information.

Research Question 4: How does the Digital Rosetta Stone Model compare with other models in relation to maintaining long-term access to digital documents? Some models deal with preservation and associated methods to ensure bit survival. Bit survival is necessary, but not sufficient to recover the stored information. Just as the ancient egyptian hieroglyphics survived, without the original Rosetta Stone they were just pretty carvings, with no other discernable information. The DRS depends on the preservation strategies being successful. It is not intended to replace them.

Research Question 5: What are the underlying assumptions of the Digital Rosetta Stone Model? Recognizing the importance of digital archaeology efforts before they are needed is inherently important. If the metaknowledge can be captured while it is available, it can be maintained for future use. If not, then the DRS's usefulness will be limited. The DRS needs to find ways to overcome the lack of critical metaknowledge. Perhaps hardware and software engineers can uncover this metaknowledge through the research of old technical and scientific journals, as well as, the U.S. Patent and Trademark Office or standards-based groups. Because the impact of the DRS is wide ranging, having buy-in from both the public and private sectors is necessary.

It is important to recognize the software that originally created a particular digital object. Unless the digital object's format is an industry standard, such as .jpeg or .bmp, that unique software's methods for interpreting the bitstream must be followed. Even if the digital object is stored using an industry standard, it is still necessary to understand the bitstream interpretation methods. While some preserved documents will be textual,

some will not. It will be necessary to understand how to read many varieties of digital objects.

Research Question 6: What steps are necessary to begin implementation of the Digital Rosetta Stone Model? It is important to know whether the DRS depends on the original medium being available at the time of data recovery. The answer is that if the medium is not available, for whatever reason, recovery efforts cannot proceed. However, the original medium need not be kept or maintained if the bitstream has been "refreshed" to a newer or technologically current medium.

Non-textual information makes up a large amount of stored digital information so it is important to know how to recover this type of non-text data. The recovery process for all data types is the same—the bits are retrieved and then interpreted. Other development and implementation questions need to be answered because at this point, the DRS is a framework, not a currently implementable solution. A consortium needs to be developed to conduct extensive research in order to build a robust solution to the long-term access problem.

Efforts to develop the DRS and MKA will be expensive and time-intensive. It is therefore necessary to know what resources will be required and when, to best manage development and implementation. Because the DRS would the link to our obsolete digital history, once developed, it needs to undergo significant testing.

Research Question 7: Who should undertake development and implementation of the Digital Rosetta Stone? And why? Involving a consortium of those who use and store information will benefit the development and implementation of the DRS. As previously

mentioned, pitfalls can be avoided and a well-designed solution crafted by those who have worked on previous access or preservation strategies, as well as others who are the beneficiaries of the DRS process.

Research Question 8: Do the experts have anything else to contribute that does not fit in the previous questions? Because the framework needs to be further developed, the DRS project needs to be brought into the contact of those, as mentioned above, to clarify and develop the DRS. If the DRS is developed in isolation, it may not be as comprehensive as it needs to be.

Non-Delphi Related Observations. The digital world is dynamic. Moore's Law (Intel, 2000a) has demonstrated that point. If there is not a concerted effort to keep track of how information is accessed and interpreted, the information that is left in obsolete media and obsolete format may be lost forever. The DRS is a strategy to address just such a problem.

Preservation of digital information is important. The need for the DRS can be reduced if the Air Force and other groups responsible for information take the appropriate steps to ensure every piece of information is kept in up-to-date form. To do this, identifying all information assets is important. Information that seems to be low value, may suddenly increase later depending on world events, so it may make sense to try to save everything, "just in case." Keeping track of everything will continue to get harder as we store more information and in greater varieties of data types.

As this group of experts has pointed out, developing the DRS is a major undertaking and expensive as well as time consuming. The focus of the DRS is on recovering

stranded data. However, data is often left behind because of its low value. Just as one uses a metal detector and invests the time to find and unearth a detected "treasure", it may turn out to be a bottle cap or rusty nail. On the other hand it could be a lost wedding band or valuable gold coin. In other words, the true value of the entire DRS effort can not be gauged until it has been developed, utilized, and the recovered data's value realized.

Limitations

This study has a few limitations. For instance, the use of the Delphi Technique does not guarantee truth. It works toward group consensus, however, expert groups are not always right. It may be discovered, in time, that other new technological solutions and/or standards better mitigate the risks that stranded data face.

As with Robertson's study (1996), this thesis does not test the technological feasibility of the DRS. It may turn out that other strategies may be more cost effective, although not necessarily providing a level of assurance that the DRS provides. However, until someone attempts to develop the MKA and build a prototype of the model, the cost aspect of building and using the DRS may not be fully appreciated.

Recommendations for Future Research

The group has agreed that the DRS warrants significant investigation at this time; and it needs to be brought into the contact of others where substantive work in this field is being done. The next step is to design and build a prototype of the DRS and demonstrate its technological feasibility with the help of software and hardware technologists.

Showing its practicable efficacy could then lead to a full-scale development and implementation of the DRS.

Summary

This thesis has examined the problem of maintaining long-term access to static digital documents using the Digital Rosetta Stone. The literature review covered several strategies for this access, but found them to be mostly preservation oriented and neglecting recovery issues. A group of experts has commented on the DRS using the Delphi Technique. These comments formed the basis for further group discussion. Overall, the group expressed concerns about the practicality of developing the DRS, but agreed that it is worthy of further study. If found to be technologically feasible and economically desirable, the DRS could well be a long-term solution to data recovery that would otherwise not be possible. The DRS is not a digital panacea though. Some data will ultimately be lost. It is the intention of the DRS design to keep that data loss to an absolute minimum.

<u>Appendix A: Letters to Potential Candidates and Organizations</u>

TO INDIVIDUAL:

Dear: __________

I am a master's student at the Air Force's Institute of Technology (AFIT - near Dayton, Ohio) and am doing my thesis research on maintaining long-term access to digital documents. It involves a group of experts commenting on a topic in rounds of discussion, in which I would like you or someone that you believe could represent your organization to participate.

The research I am doing is designed to explore the feasibility of, and add detail to, the conceptual framework of the Digital Rosetta Stone Model proposed by a previous student (Capt Steve Robertson) here at AFIT. Your anonymity will be safeguarded to permit open discussion. We will use electronic mail, and I estimate that it will take place over a period of a month. I plan to begin around the middle of August with three or four rounds of comment.

Please let me know if someone will be able to participate or not, as soon as possible, because we are looking to get this study underway in the next two weeks. As soon as I hear from you, I will send out the materials that discuss the Digital Rosetta Stone Model.

If you have any questions, please feel free to contact myself at Don.Kelley@afit.af.mil or my advisor, Dr. Alan Heminger, at Alan.Heminger@afit.af.mil. Thank you for your help and I look forward to hearing from you.

Respectfully,
Capt Kelley

DON M. KELLEY, Captain, USAF
Graduate Student
Department of Systems & Engineering Management
Air Force Institute of Technology
2950 P Street
Bldg. 640, Room 102
WPAFB, OH 45433-7765
DSN: 785-3636, x3636
COM: (937) 255-3636, x6498

TO CIO OR IT DIRECTOR:

Dear:__________

I am a master's student at the Air Force's Institute of Technology (AFIT -- near Dayton, Ohio) and I'm doing my thesis research on maintaining long-term access to digital documents. I am using a Delphi Group, which is a group of experts commenting on a topic in rounds of discussion.

This research is designed to explore the feasibility of, and add detail to, the conceptual framework of the Digital Rosetta Stone Model proposed by a previous student (Capt Steve Robertson) here at AFIT. The paper that he (and Dr. Alan Heminger, Ph. D. at AFIT) published in the journal, Communications of the AIS, is an attachment to this e-mail. The other attachment is a brief introduction of the model.

In this phase of my research, I am trying to get experts lined up. I would appreciate it if you would recommend one or two individuals in your organization to participate in this study. As experts, I would appreciate their participation and perspective that they can bring to this group. Their anonymity will be safe-guarded to permit open discussion. This discussion will take place over electronic mail; and I estimate that it will last for about one month with minimal involvement. It will begin around the middle of August and end around mid-September.

The initial round of the Delphi starts off with the experts receiving a list of questions and supporting material. Their answers and justifications (explanations) are used to develop more questions. Each successive round begins by sending out the new questions. This continues until the group reaches a consensus or little new information is added.

If you have any questions, please feel free to contact myself at Don.Kelley@afit.af.mil or my advisor, Dr. Alan Heminger, at Alan.Heminger@afit.af.mil. Thank you for your help and I look forward to hearing from you or your designees.

Respectfully,
Capt Kelley

==

DON M. KELLEY, Captain, USAF
Graduate Student
Department of Systems & Engineering Management
Air Force Institute of Technology
2950 P Street
Bldg. 640, Room 102
WPAFB, OH 45433-7765
DSN: 785-3636, x3636
COM: (937) 255-3636, x6498

Appendix B: Thank You Letter and Attachments

Dear _________,

Thank you for agreeing to participate in this research. It promises to be an exciting discussion on a very important topic. I have included, as attachments, an explanation of what this research is intended to accomplish, a brief description of the Digital Rosetta Stone Model, and the published article describing the DRS. When the group is finalized, I will begin the first round of discussion by sending (via e-mail) the first questionnaire.

If you encounter any problems or questions, please don't hesitate to e-mail me.

Thanks again for participating.

Respectfully,

Capt Kelley

Atch 1: What this research is about
Atch 2: About the DRS Model
Atch 3: Published DRS Model article

===
DON M. KELLEY, Captain, USAF
Graduate Student
Department of Systems & Engineering Management
Air Force Institute of Technology
2950 P Street
Bldg. 640, Room 102
WPAFB, OH 45433-7765
DSN: 785-3636, x3636
COM: (937) 255-3636, x6498

Appendix B Attachment 1

What this research is about...

Background

Today, digital information is being stored at rates that are astronomically high and unimaginable even a few years ago. The military, federal government, and private industry are storing thousands of gigabytes every day (Clickery, 2000) – perhaps hundreds of thousands.

As our store of digital information grows almost exponentially, the difficulties maintaining long-term access to it become increasingly large and quickly move toward infeasible. As Rothenberg (1999) explains it, it is akin to storing everything on a bed of "technological quicksand." A similar outcome occurred during the early part of the 19th century when book publishers, in trying to meet the insatiable demand for books, printed almost everything on acidic paper, which soon deteriorated (Pace, 2000).

To prevent history from repeating, we must find a strategy that will allow long-term access to information stored digitally on computing devices as they become many generations behind the current technology.

Request For Your Participation

This research seeks to determine whether the Digital Rosetta Stone is a viable strategy for mitigating the long-term access problem. This model has been developed as a framework, not a finished solution. And we need to get knowledgeable feedback on its usefulness. That's where you come in. You have been contacted because of your expertise in this area; and your unique perspective of this model is valuable. To that end,

we are requesting that you participate in a Delphi Group. This is nothing more than a group of experts who engage in discussion by answering questions (via email) on a certain topic. The discussion will take place in a series of rounds. The input from each round will be consolidated, and be used to generate questions for the next round. This is continued until a consensus is reached among the experts or little new information is added. It is expected that the rounds of discussion will take place over a period of about two months. Individual names will be removed from all comments so the discussion will be focused on the ideas, not the personalities. For this instance of the Delphi Group, the discussion will revolve around the model discussed below.

The Digital Rosetta Stone (DRS) Model

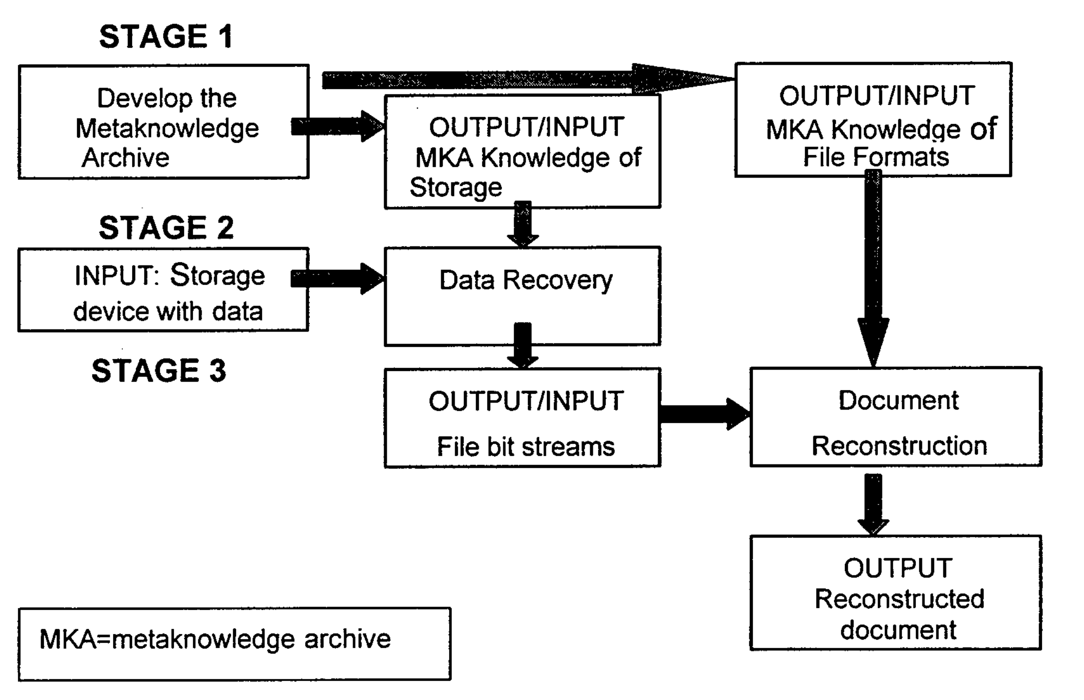


Figure 1. Digital Rosetta Stone Model

A schematic model of the DRS is shown above in Figure 1. The first stage of the model represents the knowledge preservation process. Preservation of the knowledge necessary for recovery and reconstruction of a digital document is the foundation upon which the DRS depends. During preservation the information needed to support data recovery and document reconstruction is gathered and stored in the metaknowledge archive (MKA). The types of knowledge captured in the MKA include media storage techniques and file formats. "The knowledge of media storage techniques is a collection of the way data are defined and stored on specific media ... [and] file formats is a collection of techniques used by specific software applications to define formatting operations within digital documents" (ibid).

The second stage of the model is the data recovery process. Data recovery uses the knowledge of storage techniques to extract a digital document's bit stream from an obsolete storage device and then migrates the bit stream to a currently accessible storage device. This knowledge should be of such quality to allow construction of a reader device (if no working devices exist) that could access the obsolete medium. Once a digital document's bit stream is recovered, the bit stream is advanced to the third stage.

The third stage of the model is the file reconstruction process. Document reconstruction uses the knowledge of file formats and application programs to interpret the 1's and 0's coming down the pipe and display the document in its original form. This includes all of the knowledge of how the information is bundled and other formatting concerns such as underlining and bolded items. Upon completion of the reconstruction

process, the final product is a reconstructed digital document that appears in its original form.

Why the DRS was developed

The DRS Model was developed as a conceptual model to support a strategy for maintaining access to static digital documents. Static digital documents are those that do not change over time or only have minor changes. A dynamic digital document, in contrast, is one that changes over time (perhaps often). For instance, the Internet web page for CNN is dynamic because it changes every few minutes. The DRS does not attempt to identify strategies for maintaining access to dynamic digital documents.

The DRS has been proposed in general form as a framework for capturing and maintaining the methods necessary to retrieve information. However, it has not been tested, nor have its details been worked out. It is composed of three major processes that are necessary to preserve and access our digital history -- knowledge preservation, data recovery, and document reconstruction (Robertson, 1996).

What is the Digital Rosetta Stone?

The Digital Rosetta Stone is a conceptual model for a process intended to maintain our ability to reliably retrieve and reconstruct static digital documents which are at risk for being lost because the hardware and software used to store them has become obsolete. This model, as shown in Figure 1, identifies a number of steps that, if carried out, will provide continued access to these documents. The steps include: (1)preservation of the technical knowledge of how various hardware devices and software programs store the documents, (2)using the hardware knowledge to recover the bit stream from the storage device, and (3)using the software knowledge to reconstruct the document from the bit stream.

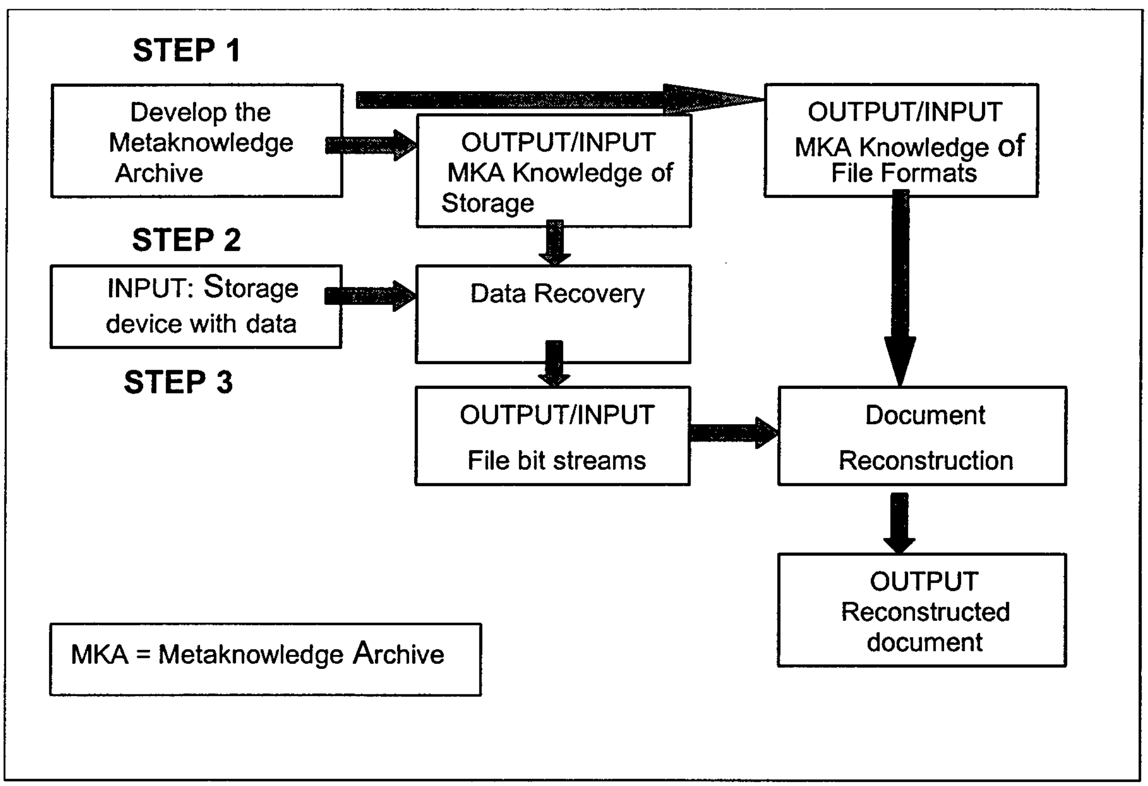


Figure 1. Digital Rosetta Stone Model

Step 1 -- Knowledge Preservation

Information is gathered on the data storage and formatting techniques used by the designers and builders of information storage devices. This includes the technical aspects of what constitutes a bit of information on this device, how it is arranged on the device, and how to access it. Information is also collected from systems and applications software that identifies the file structures, along with all information necessary to recover and read the stored digital document. The result of this step is the Metaknowledge Archive (MKA). The MKA would be developed over time by the Digital Rosetta Stone office (DRS office), and would be made available to technicians to use to recover digital documents.

Step 2 -- Bit Stream Retrieval

Information stored in the MKA would be used to access the data stored on an obsolete storage device and to transfer the bit stream to a currently accessible storage device.

Step 3 -- Document Reconstruction

The formatting information in the MKA would then be used to recover the document from the bit stream and to properly format it.

Result: The reconstructed document should be an exact replica of the original, and could then be saved on a current storage device.

Appendix B Attachment 3

For the published article titled "The Digital Rosetta Stone: A Model for Maintaining Long-Term Access to Static Digital Documents", see Communications of the Association for Information Systems Volume 3 Article 2, January 2000. The abstract is available online at http://cais.aisnet.org/articles/default.asp?vol=3&art=2.

Appendix C: Round 1 Email

All:

This e-mail is the beginning of the first round of discussion for the Digital Rosetta Stone Model. Attached are the questions for the first round. They are in a .txt and Microsoft Word 97 document format. If you have any problems or questions, please don't hesitate to e-mail me. Please be sure to respond by the 27th of August so that your thoughts can be included in the write-up for the next round. Even if you are unable to respond by that time, you will be included in following rounds.

AFIT's network administrators have been trying to make a newly installed firewall work properly, however, there have been a few e-mail outages. As such, I have set up a non-AFIT e-mail account titled DRSdelphi@aol.com. To make sure there aren't any e-mail related problems, please reply to both my don.kelley@afit.af.mil and AOL accounts.

The members of the group are

1.	David Bearman	University of Pittsburgh, School of Information Science
2.	Paul Conway	Department Head of Preservation at Yale
3.	Tim Good	Chief Information Officer of Iomega
4.	Peter Graham	Syracuse University Library Director
5.	Michael Lesk	Division Manager Computer Science Research, BellCore
6.	James Manderson	Chief, Air Force Historical Research Agency IS Division
7.	Jeff Rothenberg	Computer Scientist at RAND Corporation
8.	Don Willis	CEO of Connectex
9.	Dave MacCarn	WGBH, Chief Technologist
10.	Thom Shepard	WGBH, Universal Preservation Format Project Coordinator

Appendix C Attachment: Round 1 Questionnaire

Questions for first round of Delphi for DRS Model

Please give your opinion on each of these questions and justify your position. Again, your insight in each of these areas is greatly appreciated. Your answers and justifications for each will provide the basis for the next round of discussion. Remember to reply to both don.kelley@afit.af.mil and DRSdelphi@aol.com.

1. What are the strengths of the DRS model?
2. What are the areas in the DRS model that need improvement?
3. What is missing in the DRS model?
4. How does the DRS compare with other models for maintaining access to digital documents with which you are familiar? (Please identify the other models.)
5. What are the underlying assumptions of the DRS model? Are they valid?
6. What steps do you believe are necessary to begin implementation of the DRS model?
7. Who should undertake development and implementation of the DRS (Gov't, Industry, Consortium, other)? Why?
8. Is there anything else that you would like to address that the other questions have not asked?

Appendix D: Round 1 Responses

Of the 11 participants who agreed to respond, six actually did. Their responses to each of the questions are as follows. For reason of greater idea generation, the experts' identities will remain anonymous as far as who contributed which idea. However, to maintain some consistency as to what each expert answered, they will be identified as Experts A, B, C, D, E, and F.

Question 1: What are the strengths of the DRS Model?

Expert A: "I feel that the model is really only suitable in a very limited context. For example, a reasonable test might be to try to read a 1200 foot 800 bpi tape, or an 8 inch floppy disc. The model does not address the mechanical considerations of data recovery."

Expert B: "The model does a good job of describing the characteristics and attributes of electronic files that effect preservation and access. It lays out a methodology to maintain the ability to reliably retrieve and reconstruct digital documents."

Expert C: "I like the idea of a central registration of document types and specifications. And though I agree that capturing all this for all types that existed in the past may not be feasible, it should be possible to do this for all future types."

Expert D: "a. That it deals with digital archiving issues at all; too little attention is being paid to the problem.

b. Solutions are proposed to the category of digital archiving known as 'digital archaeology.' (See the Bill Arms model.) Digital archaeology however is regarded as the least helpful and useful mode of archiving; it assumes that no preparatory work has been done to help the end-user.

c. Dealing with what the paper calls 'cognate to paper' data. Since this category is the least difficult to archive the usefulness of the approach is necessarily limited."

Expert E: "The strengths of the Digital Rosetta Stone (DRS) are that it:
Identifies a problem of technology-locked information,
Reviews other methodologies used to preserve digital documents,

Addresses the fundamental issues of technically translating documents over time,
Evaluates general strategies for maintaining access to digital documents, and
Proposes major processes in preservation of documents."

Expert F: "It recognizes the importance of retaining a digital artifact's original characteristics by avoiding the temptation to translate its bit stream into successive new (or standardized) forms.

Unlike most approaches, it also has the potential for allowing obsolete digital storage media to be read in the future, even if no readers for such media still exist."

Question 2: What are the areas in the DRS Model that need improvement?

Expert A: "I believe the repository of metadata about the world's data is needed, but the data itself should be distributed, (Depository Libraries). Finally the mechanics of reading should consist of a device that is unlikely to change drastically, or become obsolete.......the human eye."

Expert B: "The model adequately addresses digital files that already exist but needs to provide a workable solution for the future – a standard format for document creation and markup."

Expert C: "This paper was written in 1995! There is no mention of XML or the Open Source movement and barely a discussion of digital objects.

The application that created the digital document does not have to be the one to access it. The specifications that the author writes about can be used to create new parsers (readers & browsers).

This paper has a very narrow definition of what constitutes data recovery. My zip disk may need data recovery tomorrow! This is another example why this paper errs in neglecting media degradation and media failure. The preservation problem is not merely technological obsolescence and not whether the content will be accessible in 20 years. What about digital content that is not accessible tomorrow? There are techniques used now to recovery incomplete or damaged data. Can those techniques be applied for long-term preservation?"

Expert D: "a. Extending the usefulness of the model to the areas that are of most importance, that is, of data that takes advantage of the digital environment (e.g. data base capabilities, interactive capabilities, multi-media capabilities, particular hardware that cannot easily be replicated (game-boys, early palm pilots, joysticks).

b. Cost justification for attempting this mode of preservation for 'paper cognate' materials when it is presently the wide assumption that the best mode of preservation for such materials is to print them out and save them. The gyrations gone through to do digital archaeology on old paper tape may be justified, but the future is longer than the past; why should we not simply print out paper-cognate data (in fact, mostly it is already existent in print) rather than depending on costly techniques after the fact?"

Expert E: "The improvement areas of the DRS are in the area of:
Archival relativity,
Methodology for commercial cooperation,
Functional standards for chronological interoperability,
Development of thorough metadata criteria, and
Managing documents as collective groups."

Expert F: "I see two fundamental flaws in the DRS model, one more serious than the other. The first of these is its assumption that sufficient metaknowledge about 'digital formats' can be gleaned and saved in a manner that will allow recreating the behavior of the obsolete software that originally interpreted those formats to render the documents (or other digital artifacts) they represented. I feel that the DRS inappropriately focuses on these formats rather than on the behavior of the software that interprets them: it seems to assume that the formats themselves implicitly contain all necessary information about how they are to be interpreted, which is simply not true. Much if not most of the knowledge about how digital formats are intended to be interpreted lies in the software that interprets them, not in the formats themselves. This implies that in order to work on all but the most trivial (and well-specified) of formats, the metaknowledge of the DRS would have to describe the behavior of the application programs that interpret various formats. Unfortunately, representing such behavior is an unsolved problem, and the DRS offers nothing to solve it. This is not the fault of the DRS, but its assumption that this kind of behavior can be adequately captured and represented is unwarranted. Computer science is simply not yet very adept at describing the behavior of most programs in any sort of formal way. (Describing the behavior of software informally has been proven to be inadequate time after time, if only by the failure of software development projects to realize intended requirements, which are an attempt to specify the behavior of software in advance.) Although certain aspects of the behavior of software may indeed be captured formally, much of this behavior is elusive and difficult to describe, even informally: this includes most of the 'look and feel' of software, which is an intrinsic part of the behavior of any interactive digital artifact.

The second flaw is that, despite the fact that the DRS in principle provides the ability to read obsolete digital storage media, this is generally the wrong way to approach digital preservation, except in cases of last resort. It is fundamental to digital artifacts that they consist of 'logical' bit streams, intended to be interpreted by specific programs. The physical bit stream that happens to be stored on some digital storage medium (such as a disk) is profoundly irrelevant to the artifact's logical bit stream. In light of the fact that digital documents may be stored on many different media in parallel, each having its own unique physical storage scheme, the DRS' focus on capturing metaknowledge about these physical schemes seems inappropriate. In addition, digital storage media have notoriously short physical lifetimes, beyond the fact that they become obsolete so fast. This argues that a serious, general-purpose preservation strategy should focus on preserving logical bit streams--by copying them to new storage media as necessary. The DRS' approach is somewhat unique in providing a way of recovering digital information from old, obsolete media (assuming that the information is still physically intact, which will rarely be the case); but this is hardly a general-purpose solution to preserving digital artifacts. Finally, the metaknowledge needed to read old digital media suffers from some of the same limitations described above, though the behavior of physical storage media is often much better described (and arguably easier to capture formally) than the behavior of software. Capturing such knowledge for the many different media that are in use at any given time seems a wasteful exercise, since it would be far more useful and effective to copy logical bit streams onto new media as old ones become obsolete."

Question 3: What is missing in the DRS Model?

Expert A: "The mechanics of actually making it work on a large scale. It is magnitudes more difficult to read a rotating, or moving media than a static one. We would even get read errors on some tapes when they were old, or when the read heads got dirty even though those drives were designed to read the tapes. Try reading a 2400 ft 8250bpi tape without using a tape reader designed for the job?"

Expert B: "In my opinion, the model does not address the real problem – the multitude of proprietary formats. Understand the scope of the model is to deal with the current condition of electronic documents, but the time and effort expended to develop and implement the model may be better spent on developing a long term solution."

Expert C: "There should be an explanation as to WHY institutions are storing these records digitally. What are the distinct advantages in preserving these documents as digital objects? Searching? Indexing? Transfers? Of course,

there are advantages, but the distinct properties of digital materials should be stated or reviewed here.

What also is missing is the discussion of the necessity for a "self-described" media. Specifications on how to read, for example, paper tape could have been written on the reverse side, headers could have been written in a human-readable manner, as metadata for future reconstruction.

Also, this paper infers that no progress has been made in restoring or retrieving legacy digital information. I don't think that is true.

Finally, I think this is a misguided strategy for a paper on digital preservation to omit a discussion of the instability of media?"

Expert D: "a. Primarily some sense that the authors are aware of other work going on in the field. The Australian PADI projects, the Digital Library Federation activity, the pilot projects now under way by CEDARS as part of the UK JISC effort, the Open Archiving Information Systems model created by the space science community (in which the USA is well represented), and the EC attempts to coordinate European archiving -- all these are unrepresented in the references and in the thinking displayed in the paper.

Not only does the paper show unawareness of other work, but the authors seem unaware that there could be other work. No mention is made of the archiving community or the library community, or indeed of the business and financial communities, all of which could reasonably be assumed to have an interest in dealing with this problem and to have started work on it. The indication of approaching this work in isolation is disturbing for two reasons: first, it displays a lack of organizational sophistication. Second, since the authors represent a significant military organization, it raises the possibility of a great deal of effort being put into developing a large-scale system that may satisfy military needs but will be unavailable or inadequate for other purposes. Organizational and professional intercommunication is essential in this field.

b. Analysis of cost-effectiveness of different approaches, at even a crude level. See above.

c. Any consideration of prospective data treatment, using metadata markup approaches, rather than treating all data de novo as a problem at the time user need is encountered. Some of the most effective work being done by the groups above is in developing preservation metadata approaches.

d. Recognition of the problem of authenticity, or integrity. Authenticity is the assurance we have that the information retrieved is in fact what it claims to be.

What are the checks that assure this to be the case? The obvious problems are mechanical insufficiency (dropped bits, lost segments). The DRS model can presumably take care of these. The less technical concerns are with the possibility of intentional or accidental modification of the information to serve expedient or fraudulent purposes. Systems of trust mechanisms need to be established (see Lynch et all in the CLIR publication, "Authenticity of Digital Information" (or similar title) (Council on Library and Information Resources, Wash. DC, 2000). This is a particular problem of concern if a central agency is assumed (such as the DRS agency proposed), and even more so if it is a governmental or military agency."

Expert E: "Archival distinction between a document and record
As interpreted from the paper, a document is single file of information. The impression is that documents contain their full meaning while standing on their own. However without proper context of other documents and records, the full meaning is not conveyed. The full meaning is conveyed in records and collections of related documents that are managed, stored, and maintained to preserve contextual meaning and integrity.
Example: Eighteen minutes of silence may be nothing. Eighteen minutes of silence on the ex-President Nixon tapes is significant due to the context of the gap.

Archival relativity (context, content, structure and order)
The DRS does address the structure and content of the document. It does not address the context or the order of the document in a collection. Documents may be recognized as significant at the time or over time.
Example: At the time, an E-mail document proposing a lunch date between a Lt. Colonel and a White House official may be insignificant until an independent counsel determines that the Lt. Colonel was Oliver North and the official was the White House Chief of Staff.

Selection process for document or record preservation
Many digital documents are insignificant (part of the Government administrative process). While the DRS recognizes this variability (The Digital Rosetta Stone: A model for maintaining long-term access to static digital documents – Introduction), it is unclear whether the DRS will incorporate a criteria for digital document preservation or a set of schedules for retention.

Propose the incorporation of document pedigree
A document may go through several drafts and evolve over time. It is unclear how the DRS will address the preservation of changes, modifications, and versions.

Legal issues with preservation and reconstruction
Currently, the U. S. Courts are using electronic records in the judicial system. These digital records may serve as the official document of record. Currently, there is scant case law to substantiate translation of digital official records. It is unclear how the DRS will address the legal issue of digital official records.

In addition, many vendors vehemently protect intellectual property right to their products. While the DRS mentions this point, it is unclear how the DRS will address this issue.

Multimedia preservation
The DRS covers digital textual records. It is unclear how the DRS will address multimedia and non-textual records.

Verification and Validation
Verification and validation of textual is relatively straightforward when compared to digital works of art. It is unclear how DRS will verify and validate the translation."

Expert F: "As discussed above, I feel that the model misses the importance of software in interpreting (and thereby rendering) digital documents--and the fact that the behavior of such software is not implicit in a digital artifact's logical format. In addition, I feel that the model misses the fact that it is the format of the logical bit stream of a digital artifact that is relevant to the software that renders it--and therefore to its preservation--not the physical format (or multiple formats) in which that logical format happens to be stored on particular storage media."

Question 4: How does the DRS compare with other models for maintaining access to digital documents with which you are familiar? (Please identify the other models.)

Expert A: "I proposed a model in 1992 using a combination of microfiche and computers. I sent you a copy of my paper."

Expert B: "The National Archives, San Diego Supercomputer Center, Georgia Tech Research Institute and other government agencies are working on a promising approach to store records totally independent of their hardware and software. The process is call "persistent object preservation" and uses Extensible Markup Language (XML). The approach is described in the August 28, 2000 issue of *Federal Computer Week.*"

Expert C: "Don Sawyer and Lou Reich, "Reference Model for an Open Archival Information System," White Book, Issue 4 (CCSDS 650.0-W-4.0), September 17, 1998

"SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams," Copyright (c) 1998 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc., http://www.smpte.org/engr/tfhs_out.pdf

Dave MacCarn, "Toward a Universal Data Format for the Preservation of Media," SMPTE Journal, July 1997 v106 n7 p477-479.

(Public Record Office & British Standards Institute (UK), "A Mechanism for the Perpetual Preservation of Electronic Records of Value," IDT/1/4 (A Working Group transferring to a Committee Status)TECHNICAL REPORT (Version 0.6))

These documents and others goes into specific technical requirements for a digital preservation system. They all call for the packaging or bundling of media with its metadata. The DRS needs to examine the concept of digital objects and apply its call for levels of metadata to these recommendations."

Expert D: "See above. My primary response here would be the contrast with most other models which assume that metadata creation at the time of data creation will be of the greatest use to preservation in the future."

Expert E: "Propose a review of the National Archives and Record Administration research for storage and preservation of digital records."

Expert F: "It is quite similar in spirit to the UPF (Universal Preservation Format) proposal, which grew out of a desire to preserve audio and video recordings. Both UPF and DRS rely on metaknowledge descriptions of storage formats in an attempt to allow future interpretation of those formats to reproduce originals.

It is also somewhat related to Rothenberg's proposed emulation-based approach to digital preservation (which it cites prominently), in that it recognizes the importance of retaining the original formats of digital artifacts in order to avoid corrupting them through conversion. However, it diverges from that proposal by focusing on the need to understand and formally represent logical formats in an attempt to enable future software to interpret saved digital artifacts correctly, rather than attempting to use emulation to enable running the original software that interpreted those artifacts. In addition, by focusing on preserving physical storage formats rather than

logical bit streams, it greatly complicates the problem; though this might avoid the need to copy bit streams onto new media, it does so at the cost of losing those bit streams entirely when their original storage media exceed their physical lifetimes. Finally, by focusing on the logical formats of digital artifacts, the DRS scheme would require capturing metaknowledge about hundreds if not thousands of different file formats and interpreter programs, whereas emulation requires capturing knowledge about the generally much smaller number of computing platforms on which such programs run."

Question 5: What are the underlying assumptions of the DRS model?

Expert A: "The underlying assumptions as I understand them are that all you need to recreate data written on media that has sense become obsolete is:

- knowledge preservation,
- data recovery, and
- document reconstruction.

There are so many other factors that must be considered, primarily the mechanical factors (e.g., how is the data packed, how fast does the head fly over the data, at what distance,
What is the areal density, how is the ECC incorporated, is the data stripped, what is the interface to the hardware, software, operating system. I believe the problem
to be far more complex that what has been defined.

Are they valid? They are valid, not comprehensive."

Expert B: "The assumption that specific application file format information will be available may be valid but the continued evolution of application software may result in a configuration nightmare. As Microsoft Office users, we have experienced incompatibilities between versions of the same application and found the products not as backward compatible as advertised."

Expert C: "The author's assumption is that the "native format" is what the original application generated. This seems a flawed assumption. We can't always know what application created the file, nor is it always relevant. What is considered the original application for a PDF file: Acrobat or the software the author used to create the document? Which application will allow you to view it? Ditto for html documents, as well as gifs, rtf, and postscript files. Or is the author saying that we should always save our documents in the original software's proprietary formats?
You must be able to extract the data from the original digital document in order to re-purpose it for newer applications. On the other hand, you may only need to view the document. The bottom line is that you need a system that allows you to accomplish both tasks. I believe that the trend in the computer

industry is to separate data from its different possible manifestations. Look at XML and how the same data can be presented in different ways through stylesheets."

Expert D: "a. That the data to be saved is paper-cognate. See above for the lack of usefulness, if not validity, of this assumption.

b. That data will be available on original storage media. Michael Lesk and others have made clear that preservation will mean copying (refreshing and migration) in most practical cases (always excepting Arms' "digital archaeology"). Again, it's a matter of assuming that digital preservation must always be planned for, not treated as an afterthought at a future date.

c. The Rosetta Stone model, perhaps, is itself a problematic assumption. The RS model assumes digital archaeology, which is not the situation we're in. We want to obviate the need for future Champollions, not plan for them. Unlike the Egyptians we have some sense of the finiteness of our culture and civilization (though not perhaps in an election year); unlike them we see the need to prepare for our successors, and we can do so.

d. This may not be fair as the paper is conceptual, but it seems to assume adequate resources to do whatever needs to be done: saving old media, restoring any data desired.

e. More fair may be the concern that the paper assumes that digital archiving is solely a technological problem. It is not; it is a matter of social choices. Our existing paper archives and cultural repositories exist through mechanisms determined by chance (eccentric collectors) and intentional preservation usually inadequately supported by society, requiring that triages and difficult choices be made. Digital archivists must recognize that this also will be the case, and build into the processes mechanisms for balancing need, cost and practicality.

The paper's technological emphasis is evident again in its concern for exact replication of the document. Current thinking elsewhere is sophisticated enough to understand that some "essential" quality of the data is what needs to be preserved, allowing useful (and necessarily fuzzy) arguments to take place about what is essential. Is bold facing (the document's favorite example) essential? Is tabbing and line spacing? If a multi-media document is preserved, is the resolution of the image or sound essential, and if so to what extent? Will the need be to exactly replicate an interactive document, or only to know how it conducted its interaction?

Clifford Lynch has made the useful analogy in the print world to editions. We read the Brontes in modern editions, accepting different formats (single volumes instead of three-deckers, modern typography, paperback, double-quotes in America instead of single-quotes in Britain). Only a textual scholar or an antiquarian wants to use the original "exact" text. (There is no single text of King Lear, for example; this will be the case for important digital documents; making the choice as to what to use will amenable to technological solution.)"

Expert E: "From a precursory review, the DRS assumes:
Preserved digital documents will be textual – valid,
Cooperation with the public and private sector is necessary – valid,
All digital document meaning is conveyed in bit streams – invalid,
Preserved documents meet preservation criteria – invalid,
Media metadata is rigidly defined before coming to market – invalid, and
Media metadata standards are adhered to and valid – invalid."

Expert F: "As discussed above, I see several fundamental assumptions in the DRS model that I believe to be invalid. The first is that the behavior of digital artifacts can be adequately recreated on the basis of an understanding of their logical formats, without also understanding the behavior of the original software that was intended to interpret those formats and render the artifacts. In addition, I believe that the DRS' implicit assumption that the physical formats in which the logical bit streams of digital artifacts are stored is more important than those logical bit streams is invalid.

Finally, I believe the assumption that we are capable of capturing and formally representing the necessary behavioral aspects of digital formats and the software that interprets them is unwarranted at this time--and is likely to remain so for the foreseeable future. "

Question 6: What steps do you believe are necessary to begin implementation of the DRS model?

Expert A: "I think further study of a micrographic option is warranted."

Expert B: "I assume a feasibility study has been accomplished. Before beginning, I would consider the total life cycle costs and the probability of the model being successfully implemented. I would question whether this approach is really feasible -- will the value of the information justify the expense?"

Expert C: "Preservation must be a pro-active process! The archival world should take its cue from the relative success of the Open Source initiative. The software application specifications that will most be needed (Word, et al) for this plan

to succeed will be the most difficult to obtain. The file formats, on the other hand, will be much easier. One might then deduce reader specs that will enable parsers to be built."

Expert D: "The prior question is whether it would be desirable to do so: I do not think we are ready for that decision yet. But assuming we were, the following would be necessary:

a. clarity on whether the model will attempt non-paper-cognate date, and how.

b. clarity on whether the model depends on original media being available at the time of need.

c. cost assumption explication as described above.

The model might have most use in terms of digital archaeology, but I don't think that's the most desirable place to start."

Expert E: "The DRS model is an excellent start and it is encouraging that other organizations are examining the issue of digital preservation. However, I don't believe that the DRS model is robust enough to be handle the diversity of digital information in the world today. I believe that more development needs to take place before implementation. I suggest that various consortiums be organized to further develop this model."

Expert F: "Given my reservations, I do not feel that the DRS model warrants significant investment at this time. Though I consider it a worthy goal to attempt to develop the kinds of metaknowledge that it requires, I do not believe that the necessary formalisms are likely to be forthcoming from this effort; if they are developed at all, they are more likely to come from academic research on formal computing methods and knowledge representation."

Question 7: Who should undertake development and implementation of the DRS (Gov't, Industry, Consortium, other)? Why?

Expert A: "Government through the depository library system must be given the task."

Expert B: "A consortium because of the level of involvement needed to make the model work."

Expert C: "The computer industry through consortium formation needs to take on the DRS, or some other model for the longterm preservation of digital materials. The changes & problems that occur from innovation start there. I believe that the computer industry could design "plain vanilla" application alternatives for

saving digital materials. For one example, they could offer open source "lite" versions or "reader" versions of their products. To some extent, this is already happening."

Expert D: "Implementation: this assumes that a single agency would do so; I don't make that assumption. As is the case with print (and e.g. sound recording) archiving at present, this will be a distributed activity -- and should be, for redundancy and protection against both natural and man-made disasters (war, political change, social unrest).

Development: this is a classic situation where open-source development will be of the most use and most productive. The need for the product is distributed, and multiple agencies (higher education, military, business, government, historical agencies, museums, libraries, publishers) all will have a need for interoperable systems and interchangeable archives. There is a rich tradition of standards development in these areas where interchangeable data and interoperability are desiderata, and the parties involved are accustomed to working in this tradition and do so very fruitfully."

Expert E: "Development of a model will require standards and agreement with all sectors of digital generation."

Expert F: No answer

Question 8: Is there anything else that you would like to address that the other questions have not asked?

Expert A: No answer

Expert B: "I personally do not believe this is a viable approach to long term electronic document preservation. A standardized method of marking or describing the content and relationships of information (contained within the document) which is independent of software application and operating platform, is the only cost effective and viable solution. The NARA initiative based on "persistent object preservation" using the Extensible Markup Language (XML) seems to be the most plausible approach or is at least on the right track."

Expert C: No answer

Expert D: "I apologize for seeming so negative to this point. I hope these responses will be of some use. I do think that the project as so far conceived needs to be brought into contact with others where substantive work is also going on. The

skills and sophistication of the authors should not be dissipated by going it alone."

Expert E: "What is the timetable and schedule for development of the DRS? Who is involved in the development of the DRS?"

Expert F: No answer

Report from Round 1, Clarifications, and Round 2 Questions

This document reports the analysis of the responses from the first round of the Delphi Study for the Digital Rosetta Stone. Section I contains summary statements of what I understand to be the groups' overall answers to the questions in the first round. Section II contains a brief overview of the Digital Rosetta Stone. It also addresses several of the participants' concerns regarding the model. Section III constitutes the second round of questions. There are eight topic areas corresponding to the eight questions asked in round one. The purpose of this section is to elicit participant's opinions of the statements. Thank you for your continued participation.

SECTION I – Report from Round 1

The participants all seemed to have the misconception that the Digital Rosetta Stone was a two-pronged effort, the first being preservation and the second being access. The Digital Rosetta Stone is only focused on access—it assumes preservation has already occurred. While we recognize that preservation and archiving are an extremely important area, the DRS only pertains to maintaining long-term access. The group recognized that the DRS is a major undertaking, that it sets up a central repository of metaknowledge, and that digital artifacts are important.

However, some had concerns with the design and intent of the framework. These concerns will be addressed in Section II, but as a whole, the group felt that major work needed to be done if the DRS is to be a successful venture. In part, some felt that focus

was misdirected, for example, the type of data to be rescued by the DRS was too limited or that it did not address enough issues. The only intended limitation of the model is that, as currently envisioned, it does not attempt to recover documents that require references to other, non-local, information, such as database queries or hyperlinked documents.

Some similarities were noted with other strategies. It was likened to the Universal Preservation Format and Rothenberg's emulation-based strategy. It was contrasted with the:

1. Hybrid Systems Approach
2. Persistent Object Preservation
3. Reference Model for an Open Archival Information System
4. SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams
5. Mechanism for the Perpetual Preservation of Electronic Records of Value

Most of the participants suggested that the DRS contains assumptions that are not valid. A common theme put forward was that the digital environment is too unstable for gathering the necessary information to populate the MetaKnowledge Archive. Another theme was indicative of suspicions regarding the DRS's feasibility to work properly even if the MKA was well built. Some of these concerns were based on misunderstandings of the model, which hopefully will be cleared up in Section II.

Some felt that because of the problems facing the DRS, other strategies should be pursued. Others thought that a group of people, primarily by a consortium, should develop it. The specific comments from all of the areas are itemized in Section III and are there for you to express your opinion.

SECTION II – Overview and Clarifications

General Overview of the Digital Rosetta Stone

The Digital Rosetta Stone is a framework for maintaining long-term access to static digital documents. It is designed to be used, as a last resort, not as a large-scale preservation strategy. The Metaknowledge Archive (MKA) should contain all of the information necessary to devise some method to read a storage medium and recover its bitstream. It should also contain the information necessary to format that bitstream into the original document, whether it is text, graphics, video, or other. If the format does not implicitly contain all the necessary information to properly format the bitstream, then that additional information should also be in the MKA. The MKA is not intended to be used to develop an exact replica of the original software nor provide the same functionality. Once the original digital object has been reconstructed, the goal of the Digital Rosetta Stone has been fulfilled. It is not concerned with what happens to the information after recovery.

Concerns addressed here: (This is an attempt to clarify the model and clear up misunderstandings about the model.)

Topic: <u>Areas of the Model that need improvement</u>

Concern

"The Digital Rosetta Stone applies only to text-based material."

Clarification

The design of the Digital Rosetta Stone Model, is intended to work for any medium and any format. It is not limited to text documents, paper-like materials, or still image objects. Although the likelihood of capturing every hardware and software specification adequately may not be met, it is the intent for as much information to be gathered for the Metaknowledge Archive as possible. The cost of developing and implementing the DRS is better justified because it covers significantly more that text-like materials. Printing out information may be a good way to maintain access to that kind of information, but is not a strategy without serious investments of resources and management, especially when petabytes of information are considered, as well as legal issues for items "born digitally". A recent study out of California, Byte Counters, by Peter Lyman and Hal Varian, concluded that we are currently storing about 1.5 exabytes of information annually. (That's $1.5\text{x}10^{18}$ bytes.)

Concern

"The DRS does not address the mechanical considerations of data recovery."

Clarification

The DRS specifically addresses the mechanical considerations necessary for data recovery. That information would be contained in the Metaknowledge Archive. The first part of document, or digital object, reconstruction is concerned with being able to read the medium and recover the bitstream. Unlike most approaches, it has the potential for allowing obsolete digital storage media to be read, even if no readers for such media still exist. Without that technological information, it would be impossible to construct a

viable reader. The MKA would contain information such as how close the head must be to the medium, how the information is physically stored on the medium (tracks, sectors, cylinders, etc), and how to build a device to read the medium, along with the formatting information explaining all of the codes to give the digital object its original "look and feel".

There is a problem when trying to decipher what version of a format a particular recovered digital object is in. Many versions exist for the same file extension, such as .doc for a Microsoft Word document. There are even other vendors who used a .doc extension, such as WordPerfect's creator. As far as different formats for the same document extension or for those without an extension, the DRS would need to do be able to understand the different versions and be able to perform a brute force attempt to see if a digital object made sense when formatted according to different versions. In some instances, either when the MKA has not captured all of the information necessary to properly format the bitstream or when the digital object does not contain enough information to describe itself, a poor rendition of the original object may result. It may not be necessary to know what software "created" the data. In the case of Adobe's .pdf format, its very nature is portability (hence, pdf - portable document format). Many software applications can read a pdf file and it is not necessary, indeed, it may not be possible to know which application actually created the data. All that is needed is to know what Adobe's formatting standards were and how to reconstruct pdf files. This applies to many other formats such as .gif, .jpg, .jpeg, .html, etc.

Concern

There are better methods of preservation than trying to build something from scratch.

Clarification

The DRS does not supersede preservation strategies. On the contrary, it can be used in conjunction with them. For instance, a user could reconstruct a digital object and then migrate it to a newer medium and format. It is not intended to be a preservation strategy.

Concern

"It needs to provide a workable solution for the future—a standard format for document creation and markup."

Clarification

The model is not designed to be a standard format for preservation, as in the Universal Preservation Format or other similar strategy. It is not a preservation strategy, it is one of recovery. It is designed to be the last attempt when all else has failed. The existence of a few formats, or only one, would make it significantly easier to reconstruct a digital object. This is not currently the case, however, and is not likely to be in the near term. Even if such a universal format existed tomorrow, there currently exists 40 to 50 years worth of distinctly different formats of both hardware and software.

Concern

"There is no mention of XML or the Open Source movement and barely a discussion of digital objects."

Clarification

Even though there is little mentioned about digital objects and nothing about the eXtensible Markup Language (XML), the Digital Rosetta Stone, as it exists today is a framework or a concept. The details of design and implementation as well as public, private, and foreign cooperation all needs to be further developed. The lack of detail should not limit its possibility of being a tremendously successful endeavor--only that we should work more to see it through.

We also need to develop the MKA criteria and its format. Currently, the Air Force Institute of Technology, creator of the DRS, does not have a timetable to develop the DRS.

Concern

"What about digital content that is not accessible tomorrow?"

Clarification

As mentioned earlier, the Digital Rosetta Stone is designed to be a last ditch effort in recovering information. While it is acknowledged that this is regarded as the least helpful and useful mode of archiving, it is not designed to be an archiving strategy. There are other strategies that should be used to convert or preserve large quantities of data on a large scale that is not on obsolete equipment. It would not make sense to try to build a storage medium reader and develop software when working instances currently exist, unless it is for testing the DRS as a viable solution. It is, however, expected that the managers, owners, and creators of the information should do what is necessary to make sure that the information will remain accessible far into the future. While such a scenario

would be ideal, we assume that because of the tremendous amount of information that currently exists and the fact that it resides in facilities that have fundamentally different access and preservation requirements, some information will be left behind. It is precisely this stranded information that the DRS seeks to recover. As an access strategy, it assumes that the media have been physically preserved. It does not prescribe specifications for a physical storage environment or anything else that applies to preservation. Also, because the field of digital preservation is so new and we lack a long-term solution, we face data loss every day. In fact, we have already irretrievably lost data (1960's US Census, NY State hazardous site locations, and a plethora of other instances). This loss of information is due to media degradation and mishandling. While there have not been any major reports of data loss due to loss of access knowledge, there have been a few close calls. Therefore, we need to develop the recovery strategies before they are needed.

There are some specialized techniques that are currently used to recover data on media that have been physically damaged. The techniques, if well documented, could be useful for digital object reconstruction. This material should be captured in the Metaknowledge Archive.

It is assumed that some recovered information is better than no information. Therefore, even if the MKA does not contain all of the information that is necessary to provide the original "look and feel" of a digital object, some data may be recovered to provide a degraded, but somewhat helpful digital object. There are some serious challenges to overcome in collecting the information that will constitute the MKA.

Among these are that most of it is proprietary information that is closely guarded, it is sometimes not well defined, and standards are not always followed.

Topic: What is missing in the Model?

Concern

Selection processes, retention schedules, media preservation, and verification and validation procedures are all missing.

Clarification

The DRS is designed to be a recovery tool, not as a silver bullet to manage documents. It is expected that institutions concerned with digital document access such as libraries and digital archives will have some sort of document management procedures in place. While we recognize that digital archiving is more than just a technological problem, the DRS is focused on the digital document recovery process. The other issues such as social choices and the legal arena are not directly involved with the physical act of document recovery.

Topic: What is the next step?

Concern

"The model might have some use in terms of digital archaeology, but I don't think that's the most desirable place to start."

Clarification

Digital archaeology is the focus of the DRS even though it might not be where one would want to start out from. However, history has shown us that it is the unfortunate position that we are in. The DRS seeks to overcome the problems typically associated with digital archaeology, such as finding an unrecognized medium and not knowing where in the world to start.

SECTION III – Round 2 Questions

Instructions: Each of the items below was listed by one of the participants. The first round concerns that have been clarified are not included. Please indicate how much you agree with the statements. Also answer how relevant the statements are as they apply to the Digital Rosetta Stone and maintaining long-term access to static digital documents, rather than as a preservation strategy. The number in parentheses at the end of the statement is the number of participants who submitted it. It may be helpful to review the DRS article and other papers from the first round. Thank you for your continued participation.

Please indicate your agreement level in the first column.

Indicate in the second column how important this topic is to the model's usefulness as a strategy for maintaining long-term access to static digital documents

1	2	3	4	5
Totally Disagree	←	**Agreement**	→	Totally Agree

1	2	3	4	5
Not Important	←	**Importance**	→	Very Important

Question 1: What are the strengths of the DRS model ?	**Agreement**	**Importance**
1. It recognizes the importance of retaining access to objects even as the technology for storing them becomes obsolete.(4)		
2. It recognizes the importance of the digital object's original characteristics.(2)		
3. It allows for access even if no readers for such a medium exists.(1)		
4. It lays out a methodology to maintain the ability to reliably retrieve and reconstruct digital documents.(1)		
5. It has the idea of a central registration of document types and specifications.(1)		
6. It addresses fundamental issues of technically translating documents over time.(1)		
Additional comments:		

1 Totally Disagree	2	3 Agreement	4	5 Totally Agree

1 Not Important	2	3 Importance	4	5 Very Important

Question 2: What are the areas in the DRS model that need improvement?	**Agreement**	**Importance**
1. The Metaknowledge Archive should have its data distributed instead of centralized.(1)		
2. The DRS has too narrow a view of what constitutes data recovery, i.e., it should include a short-term perspective as well.(1)		
3. It doesn't describe how to handle media degradation and media failure.(1)		
4. Where possible, the DRS should integrate well with archiving.(2)		
5. It needs to spell out a methodology for commercial cooperation.(1)		
6. It needs to develop functional standards for chronological interoperability.(1)		
7. The Metaknowledge criteria needs to be further developed.(1)		
8. The DRS should focus more on the behavior of the software that interprets the bitstream rather than on the format of the physical medium.(1)		
Additional comments:		

1	2	3	4	5
Totally Disagree		Agreement		Totally Agree

1	2	3	4	5
Not Important		Importance		Very Important

Question 3: What is missing in the DRS model?	**Agreement**	**Importance**
1. The need for self-describing metadata.(1)		
2. It doesn't address media instability.(1)		
3. The awareness of other long-term access efforts and its compatibility with them.(1)		
4. An analysis of the cost-effectiveness of different approaches.(1)		
5. It doesn't address what to do with the data after recovery.(1)		
6. It doesn't address the problem of authenticity, or integrity, of the original document.(1)		
7. It lacks the archival distinction between a document and a record.(1)		
8. It does not address the context or order of the document in a collection.(1)		
9. It doesn't address any legal-related issues such as intellectual property rights.(1)		
10. It doesn't address verification and validation of the translation.(1)		
11. It misses the importance that software plays in interpreting the digital documents by the fact that the behavior of such software is not implicit in a digital artifact's format.(1)		
12. It misses the fact that it is the format of the logical bitstream that is important to the software and presentation of the data -- not the implementation of how it is physically stored on a medium.(1)		
Additional comments:		

1	2	3	4	5
Totally Disagree	←	Agreement	→	Totally Agree

1	2	3	4	5
Not Important	←	Importance	→	Very Important

Question 4: How does the DRS compare with other models for maintaining access to digital documents with which you are familiar? (Please identify the other models.)	**Agreement**	**Importance**
1. It differs from a hybrid systems approach to preservation of printed materials by Don Willis in that it is a strategy for long-term access instead of preservation.(1)		
2. It differs from Persistent Object Preservation by the fact that it is a method for maintaining long-term access instead of a method of preservation.(1)		
3. Other schemas are geared toward digital document preservation.(2)		
4. The Digital Rosetta Stone is very similar to the Universal Preservation Format.(1)		
5. The DRS is related to Rothenberg's emulation-based strategy in that it recognizes the importance of retaining the original formats. However, it diverges in the fact that the emulators are used to properly interpret the bitstream.(1)		
Additional comments:		

1 Totally Disagree	2	3 Agreement	4	5 Totally Agree

1 Not Important	2	3 Importance	4	5 Very Important

Question 5: What are the underlying assumptions of the DRS model?	**Agreement**	**Importance**	**Is this assumption valid? (Y/N)**
1. All that is needed to recreate data written on media are knowledge preservation, data recovery, and document reconstruction.(1)			
2. The Metaknowledge Archive will be available.(1)			
3. The "native format" is what the original application created.(2)			
4. Data will be available about the original storage media.(1)			
5. The DRS assumes we are in a situation that needs digital archaeology.(1)			
6. Assumes adequate resources will be provided.(1)			
7. DRS assumes digital archiving is solely a technological problem.(1)			
8. Some preserved digital documents will be textual.(1)			
9. Cooperation with the public and private sectors is necessary.(1)			
10. All of a digital document's meaning is conveyed in bitstreams.(1)			
11. Preserved documents meet preservation criteria.(2)			
12. Media metaknowledge is rigidly defined before coming to market.(1)			
13. Media metaknowledge standards are valid and adhered to.(1)			
14. The physical formats in which the logical bitstreams of digital artifacts are stored is more important than the logical bitstreams.(1)			
Additional comments:			

1 Totally Disagree	2	3 Agreement	4	5 Totally Agree

1 Not Important	2	3 Importance	4	5 Very Important

Question 6: What steps do you believe are necessary to begin implementation of the DRS model?	**Agreement**	**Importance**
1. Assuming a feasibility study has been performed, consider the total life cycle costs and probability of the model being successfully implemented.(2)		
2. Accumulate the metaknowledge.(1)		
3. Assuming we are ready for a decision, clarify how the model would attempt to recover non-textual information.(1)		
4. Clarify whether the model depends on the original medium being available at the time of need.(1)		
5. Development of the consortium to further build the model.(1)		
6. The DRS does not warrant significant investigation at this time.(1)		
Additional comments:		

Question 7: Who should undertake development and implementation of the DRS (Gov't, Industry, Consortium, other)? Why?	**Agreement**	**Importance**
1. Government through the depository library system.(1)		
2. A consortium of those who store and use information.(4)		
Additional comments:		

1	2	3	4	5
Totally Disagree	←	**Agreement**	→	Totally Agree

1	2	3	4	5
Not Important	←	**Importance**	→	Very Important

Question 8: Is there anything else that you would like to address that the other questions have not asked?	**Agreement**	**Importance**
1. This project needs to be brought into the contact of others where substantive work in this field is being done.(1)		
Additional comments:		

Thank you for your time in participating in the second round. Please take a few minutes to look over your answers and make sure they are categorized according to 1 for Totally Disagree, 5 for Totally Agree on the Agreement Column and 1 for Not Important, 5 for Very Important on the Importance Column. Once you are finished, please send your answers to both DRSdelphi@aol.com and don.kelley@afit.af.mil.

Appendix F: Round 2 Data and Statement Categorization

		Statement Number	Importance 1	2	3	4	5	median	Agreement 1	2	3	4	5	median	Validity
Important	Agree	1.1	1	0	0	1	6	5.0	0	0	1	3	4	4.5	
		1.3	0	0	1	1	6	5.0	2	1	0	3	2	4.0	
		1.5	0	3	1	2	2	3.5	0	1	1	4	2	4.0	
		2.3	2	0	1	2	2	4.0	1	1	1	4	0	4.0	
		2.4	0	0	0	4	3	4.0	0	1	0	4	2	4.0	
		2.7	0	2	1	3	2	4.0	0	0	2	3	2	4.0	
		3.1	0	1	1	3	3	4.0	0	2	1	2	2	4.0	
		3.3	0	0	2	4	2	4.0	1	0	1	2	3	4.0	
		3.6	1	1	0	3	3	4.0	1	2	0	3	1	4.0	
		3.10	0	1	1	4	2	4.0	0	1	0	4	2	4.0	
		3.11	1	0	2	4	1	4.0	0	2	0	4	1	4.0	
		4.3	1	0	2	3	0	3.5	0	0	3	3	0	3.5	
		5.3	1	1	0	5	0	4.0	0	2	0	3	2	4.0	Valid
		5.5	0	2	1	2	1	3.5	2	0	1	2	2	4.0	Valid
		5.8	1	0	1	1	3	4.5	0	0	0	1	6	5.0	Valid
		5.9	0	1	1	1	4	5.0	0	0	1	1	5	5.0	Valid
		6.1	1	0	2	1	3	4.0	1	0	1	2	3	4.0	
		6.3	0	1	0	5	1	4.0	0	1	0	6	0	4.0	
		6.4	1	0	2	3	1	4.0	1	0	1	3	2	4.0	
		6.5	0	1	0	3	3	4.0	0	1	0	2	4	5.0	
		7.2	0	0	2	3	2	4.0	0	0	1	3	3	4.0	
		8.1	0	1	0	2	2	4.0	0	0	2	1	3	4.5	
	Unsure Agree	1.6	0	0	0	2	6	5.0	2	2	0	3	1	3.0	
		3.2	1	2	1	3	1	3.5	1	2	1	2	1	3.0	
		3.12	0	0	1	4	3	4.0	1	1	2	2	1	3.0	
		5.4	1	0	1	3	2	4.0	1	2	1	2	1	3.0	Unsure
		5.14	0	1	2	2	1	3.5	0	2	3	1	0	3.0	Invalid
	Disagree	1.4	0	0	0	1	7	5.0	2	2	1	2	1	2.5	
		2.8	1	0	3	3	1	3.5	1	3	0	2	1	2.0	
		5.1	0	0	1	4	3	4.0	3	3	1	1	0	2.0	Invalid
		5.2	0	1	0	1	4	5.0	2	1	1	2	0	2.5	Unsure
		5.6	0	0	0	2	4	5.0	2	3	1	1	0	2.0	Invalid
		5.10	0	0	2	5	0	4.0	2	3	0	1	1	2.0	Invalid
		5.11	0	1	1	3	1	4.0	0	5	1	0	0	2.0	Invalid
		5.13	0	2	1	2	1	3.5	1	3	2	0	0	2.0	Invalid
		6.6	0	0	2	3	2	4.0	1	4	1	1	1	2.0	
Unsure Important	Agree	1.2	0	2	3	1	2	3.0	0	0	1	3	4	4.5	
		7.1	1	0	3	3	0	3.0	1	1	1	4	0	4.0	
		6.2	0	0	4	2	1	3.0	0	1	2	4	0	4.0	
		4.5	0	0	4	3	0	3.0	0	2	1	5	0	4.0	
		2.5	0	1	4	1	2	3.0	0	1	1	4	1	4.0	
		4.2	1	1	2	1	1	3.0	0	0	3	1	2	3.5	
		3.4	0	1	3	3	0	3.0	1	1	1	2	1	3.5	
	Unsure Agree	4.4	0	0	4	2	0	3.0	0	0	5	2	0	3.0	
		4.1	1	0	3	2	0	3.0	0	1	3	1	1	3.0	
		2.6	1	1	3	2	0	3.0	0	2	2	2	0	3.0	
		2.1	1	1	3	2	1	3.0	1	1	2	1	2	3.0	
	Disagree	5.12	0	2	2	1	1	3.0	1	2	3	0	0	2.5	Invalid
		5.7	1	0	3	2	1	3.0	2	3	1	1	1	2.0	Invalid
Not Important	Agree	3.5	3	2	1	1	1	2.0	2	1	0	4	0	4.0	
		3.8	3	2	0	2	0	2.0	1	1	1	1	2	3.5	
	Unsure Agree	3.7	3	1	1	0	2	2.0	1	1	1	0	2	3.0	
		3.9	3	1	1	3	0	2.5	1	1	2	1	2	3.0	
	Disagree	2.2	3	2	2	1	0	2.0	1	3	2	1	0	2.0	

Important	Agree	1.1 It recognizes the importance of retaining access to objects even as the technology for storing them becomes obsolete.(4)
		1.3 It allows for access even if no readers for such a medium exists.(1)
		1.5 It has the idea of a central registration of document types and specifications.(1)
		2.3 It doesn't describe how to handle media degradation and media failure.(1)
		2.4 Where possible, the DRS should integrate well with archiving.(2)
		2.7 The Metaknowledge criteria needs to be further developed.(1)
		3.1 The need for self-describing metadata.(1)
		3.3 The awareness of other long-term access efforts and its compatibility with them.(1)
		3.6 It doesn't address the problem of authenticity, or integrity, of the original document.(1)
		3.10 It doesn't address verification and validation of the translation.(1)
		3.11 It misses the importance that software plays in interpreting the digital documents by the fact that the behavior of such software is not implicit in a digital artifact's format.(1)
		4.3 Other schemas are geared toward digital document preservation.(2)
		5.3 The "native format" is what the original application created.(2)
		5.5 The DRS assumes we are in a situation that needs digital archaeology.(1)
		5.8 Some preserved digital documents will be textual.(1)
		5.9 Cooperation with the public and private sectors is necessary.(1)
		6.1 Assuming a feasibility study has been performed, consider the total life cycle costs and probability of the model being successfully implemented.(2)
		6.3 Assuming we are ready for a decision, clarify how the model would attempt to recover non-textual information.(1)
		6.4 Clarify whether the model depends on the original medium being available at the time of need.(1)
		6.5 Development of the consortium to further build the model.(1)
		7.2 A consortium of those who store and use information. (4)
		8.1 This project needs to be brought into the contact of others where substantive work in this field is being done.(1)
	Unsure Agree	1.6 It addresses fundamental issues of technically translating documents over time.(1)
		3.2 It doesn't address media instability.(1)
		3.12 It misses the fact that it is the format of the logical bitstream that is important to the software and presentation of the data -- not the implementation of how it is physically stored on a medium.(1)
		5.4 Data will be available about the original storage media.(1)
		5.14 The physical formats in which the logical bitstreams of digital artifacts are stored is more important than the logical bitstreams.(1)
	Disagree	1.4 It lays out a methodology to maintain the ability to reliably retrieve and reconstruct digital documents.(1)
		2.8 The DRS should focus more on the behavior of the software that interprets the bitstream rather than on the format of the physical medium.(1)
		5.1 All that is needed to recreate data written on media are knowledge preservation, data recovery, and document reconstruction.(1)
		5.2 The Metaknowledge Archive will be available.(1)
		5.6 Assumes adequate resources will be provided.(1)
		5.10 All of a digital document's meaning is conveyed in bitstreams.(1)
		5.11 Preserved documents meet preservation criteria.(2)
		5.13 Media metaknowledge standards are valid and adhered to.(1)
		6.6 The DRS does not warrant significant investigation at this time.(1)
Unsure Important	Agree	1.2 It recognizes the importance of the digital object's original characteristics.(2)
		7.1 Government through the depository library system.(1)
		6.2 Accumulate the metaknowledge.(1)
		4.5 The DRS is related to Rothenberg's emulation-based strategy in that it recognizes the importance of retaining the original formats. However, it diverges in the fact that the emulators are used to properly interpret the bitstream.(1)
		2.5 It needs to spell out a methodology for commercial cooperation.(1)
		4.2 It differs from Persistent Object Preservation by the fact that it is a method for maintaining long-term access instead of a method of preservation.(1)
		3.4 An analysis of the cost-effectiveness of different approaches.(1)
	Unsure Agree	4.4 The Digital Rosetta Stone is very similar to the Universal Preservation Format.(1)
		4.1 It differs from a hybrid systems approach to preservation of printed materials by Don Willis in that it is a strategy for long-term access instead of preservation.(1)
		2.6 It needs to develop functional standards for chronological interoperability.(1)
		2.1 The Metaknowledge Archive should have its data distributed instead of centralized.(1)
	Disagree	5.12 Media metaknowledge is rigidly defined before coming to market.(1)
		5.7 DRS assumes digital archiving is solely a technological problem.(1)
Not Important	Agree	3.5 It doesn't address what to do with the data after recovery.(1)
		3.8 It does not address the context or order of the document in a collection.(1)
	Unsure Agree	3.7 It lacks the archival distinction between a document and a record.(1)
		3.9 It doesn't address any legal-related issues such as intellectual property rights.(1)
	Disagree	2.2 The DRS has too narrow a view of what constitutes data recovery, i.e., it should include a short-term perspective as well.(1)

Appendix G: Round 2 Report

Report from Round 2

Introduction

These findings were gathered from the expert's responses. I categorized the items into eight areas or topics that each statement seemed to address. They are ordered in a manner that tries to present an overall picture of the DRS landscape. A matrix of categories for opinions was developed. This facilitated categorization of each of the statements based on the level of consensus on statement importance and statement agreement.

Statement Topics

The first topic deals with the preservation and access environment that created the need for the DRS. The second topic deals with physical media devices and digital objects. Given this environment in which we find ourselves, the third topic covers relevant areas of the development of the Digital Rosetta Stone. The fourth covers the focus of the DRS. The fifth topic covers the methodology of the DRS and the following two areas, six and seven, go into more detail of the methodology category. The eighth, and last, area deals with statements made about the DRS implementation details. These topics are designed to give the reader some idea about where each of the statements belong in the DRS landscape.

Statement Topics

1. Preservation and Access Environment
2. Media and Digital Objects
3. Development of the DRS
4. DRS Focus
5. DRS Methodology
6. Metaknowledge Archive
7. Software, Logical Formats and Physical Formats
8. DRS Implementation Details

The experts submitted opinions about the statements in the form of two parts. The first opinion was directly related to whether or not the expert agreed with the statement. The second opinion dealt with whether or not the statement was important to the DRS. The opinions were recorded using a 5-point Likert-type scale, with the low end being either disagree or not important. High numbers were used to indicate agreement or high importance. Question 5 related to assumptions that the DRS made. The experts were also asked to state if these assumptions regarding Question 5 were valid or not.

Each of the statements has two opinion parts: statement agreement and statement importance. Each of the opinion parts has three possible answers: Agree/Important, Unsure/Unsure, or Disagree/Unimportant. This results in nine possible statement agreement and statement importance opinion outcomes or categories.

		Levels of Importance		
		High (A)	Unsure (B)	Low (C)
Levels of Agreement	High (1)	Important and Agree	Unsure Important and Agree	Not Important and Agree
	Unsure (2)	Important and Unsure Agree	Unsure Important and Unsure Agree	Not Important and Unsure Agree
	Low (3)	Important and Disagree	Unsure Important and Disagree	Not Important and Disagree

Figure 1. Categories for Opinions

For purposes of tracking which statements belong in what category, each row and column has been labeled with a letter or number, in addition to the level of importance or agreement. The Importance Level columns have been labeled A, B, and C, corresponding to their order. The Agreement Level rows have been labeled with 1, 2, and 3. For example, the category of Important and Agree will be referenced as Category A1. The Important and Disagree category will be referred to as Category A3. Also, each one of the eight statement topics will be referred to by its corresponding number. Every opinion discussed in this report will have a similar heading consisting of the category rating (A1, A2, A3, B1, etc.) and statement topic number (1-8). In the case of the first opinion, the heading will be "A1.1 Preservation and Access Environment"—A1 being the category for the Important and Agree opinions.

Not every one of the nine categories for opinions had every statement topic in it, but all of the topics fit into the categories. The statement topics will be discussed by level of importance followed by level of agreement.

The Difference Between a Group Rating of Unsure and Disagree or Not Important

There is a fine distinction that needs to be made between a rating of Unsure and a rating of Disagree or Not Important. For instance, the group could come to a consensus on a statement--deciding that it was important but disagree with it. This disagreement should not be confused with not having a consensus. If all of the experts said they disagreed with a statement, then the group would have come to a consensus that they, as a whole, disagreed with a statement. The points where the group did not come to a

consensus, either for importance or agreement, are listed as Unsure. Also, a statement could be listed in the Unsure category based on a group consensus of unsure.

Discussion of the Group's Opinions on Each of the Statements

A1. Important and Agree Category

This category consists of those topics on which the group of experts reached a consensus that they agree with the statements and also agree that the statements were materially important to the Digital Rosetta Stone and its development. One third of the statements fell in this category.

A1.1 Preservation and Access Environment

As young as the digital world is, we are already seeing that there is a definite need for digital archaeology. This validates the DRS assumption of a need for digital archaeology. If the DRS is to be successful, it needs to be aware of other strategies for long-term access and those for preservation as well as be compatible with them.

A1.2 Media and Digital Objects

Making sure that the output matches the original is important. The developers of the DRS need to take this into account. Because of the long-term nature of the DRS and the general instability of media, the DRS should seek to use or develop methods to handle media degradation and failure. To aid in future recovery efforts, the developers should address the need for self-describing media, although the DRS does not currently do this. To the extent that this could be done, utilizing self-describing media would certainly

simplify the DRS. It would assist in the process of recovering the bitstream, leaving only the interpretation of the bitstream to complete document recovery.

A1.3-5 No statements in these topics fell in the A1 category.

A1.6 Metaknowledge Archive

The DRS can accomplish its long-term access mission because it maintains the Metaknowledge Archive. Because the foundation of the DRS is the MKA, the criteria for the MKA needs to be developed further and clearly specified.

A1.7 Software, Logical Formats and Physical Formats

Software is very important, and a concerted effort with software developers will be necessary to capture sufficient information to assist the DRS. Some files are application independent, such as .jpeg or .bmp. The "native format" is the format that the originating software used for the file and this format is important to understand. Some of these digital documents will be textual or paper like, but the rest will not. The DRS needs to clarify how the model would attempt to recover the non-textual digital objects. These digital information object types could include anything from database files to graphics to encapsulated metadata digital objects.

A1.8 DRS Implementation Details

The group strongly agrees that maintaining long-term access to documents is important and that the DRS allows for that access even if no readers exist for that medium. The sentiment was not unanimous—there was one who disagreed on the DRS portion of the statement. Cooperation for implementing the DRS with the public and private sectors is necessary. The development process should include a prototype to

determine technical feasibility, total life-cycle cost analysis, and a probability determination of a successful DRS implementation. A consortium of those who store and use information needs to be developed to further build the model. To help get the process of DRS development going, it needs to be exposed to others where substantive work is being done in this field.

A2. Important but Unsure of Agreement Category

These issues are important to the DRS but the experts are not sure if they agree with the items or not.

A2.1 Preservation and Access Environment

Addressing the fundamental issues of technically translating documents over time is important, but the experts are unsure that the DRS does this. At this point in its infancy, the DRS does not yet actually cover the technical issues; it will do this when the model is developed.

A2.2 Media and Digital Objects

Media instability is an important problem, but the group is unsure if the DRS addresses that problem. Data about the original storage media are important, but the group is unsure that the data will be available when it comes time to capture it for the MKA. The group is not sure that the DRS make the assumption that this data will be available. One expert says that it is easier to capture the data when it is readily available.

A2.3-6 No statements in these topics fell in the A2 category.

A2.7 Software, Logical Formats and Physical Formats

The format of the logical bitstream is important to the software and how the data is presented is important, but the group is unsure whether or not the DRS misses this point.

The group is not sure if the DRS makes the assumption that the physical format of the digital artifact's logical bitstream is more important than the logical bitstream itself. They do not think that the physical format is more important than the logical bitstream. In other words, both the physical format and the software formats are important to data recovery.

A2.8 This statement topic did not fall in the A2 category.

A3. Important and Disagree Category

This grouping of items was found to be important to the DRS, but the experts disagreed with the statements. This suggests a consistency in responses, because some of the statements were relatively opposite with what some of the agree statements were.

A3.1-4 No statements in these topics fell in the A3 category.

A3.5 DRS Methodology

The group thinks a methodology to maintain the ability to reliably retrieve and reconstruct digital documents is important but they do not think that the DRS has such a methodology. It could be that they do not think it does yet or that it will not have one at all. I would agree at this point, the methodology is not fully developed. They agree that adequate resources are necessary but do not think that the DRS assumes that the needed

resources will be available. The DRS is important and does warrant significant investigation at this time.

A3.6 Metaknowledge Archive

The group agrees that the MKA is important but is unsure if the MKA will be available. They do not think that the DRS makes this assumption. Preserved documents are important but do not necessarily meet preservation criteria. The group does not think the DRS makes this assumption either. Media metaknowledge standards are important, but are not adhered to or valid. The group does not think the DRS makes this assumption.

A3.7 Software, Logical Formats and Physical Formats

The group thinks that software behavior and physical format are important but that the DRS should not focus more on the software behavior than the physical format. Data re-creation is important but knowledge preservation, data recovery, and document reconstruction are not all that is needed. They also do not think that the DRS makes this assumption. A digital document's meaning is important but not entirely conveyed by the bitstream. They agree that the DRS does not make this assumption.

A3.8 No statements in this topic fell in the A3 category.

B1. Unsure Important and Agree Category

The experts were unsure of how important these items were to the DRS but did reach a consensus on agreement for each item.

B1.1 Preservation and Access Environment

The DRS is related to Rothenberg's emulation-based strategy in that it recognizes the importance of retaining original formats. They also agree that it diverges in the fact that the emulators are used in Rothenberg's solution to properly interpret the bitstream, but not in the DRS. They are not sure how important this statement is to the DRS.

It differs from Persistent Object Preservation because the DRS is an access method not a preservation method. Because it does differ, the group is unclear on how important Persistent Object Preservation is in terms of impact on the DRS.

B1.2 No statements in this topic fell in the B1 category.

B1.3 Development of the DRS

They agree that the government should help undertake the implementation of the DRS but are not sure how important or to what level the government should have in its involvement.

B1.4 DRS Focus

The group agrees that the DRS recognizes the importance of the digital object's original characteristics, but rates the importance as "unsure".

B1.5 DRS Methodology

The DRS needs to spell out a methodology for commercial cooperation, but the group is unsure how important it is to the overall success of the DRS. They agree that it needs to have an analysis of the cost-effectiveness of other approaches. This goes to the overall awareness of the other methods as stated previously.

B1.6 Metaknowledge Archive

The metaknowledge should be accumulated, however, the group is unsure how this will affect the overall implementation of the DRS.

B1.7-8 No statements in these topics fell in the B1 category.

B2. Unsure Important and Unsure Agree Category

The group was unsure of how important these items are to the DRS and are ambivalent about whether or not the group agrees with these statements.

B2.1 Preservation and Access Environment

The group was unsure of how the DRS compared to a hybrid systems approach for preservation of printed materials and was also not sure how this applied to the DRS. The group was unsure of whether the DRS was similar to the Universal Preservation Format. This is not surprising because the experts may not have been familiar with the UPF.

B2.2-4 No statements in these topics fell in the B2 category.

B2.5 DRS Methodology

The group was unsure of whether the MKA should be distributed or centralized. They were also unsure of how important the level of centralization or decentralization was to the DRS. They were unsure of whether it needed to develop functional standards for chronological interoperability. They were also unsure of how important this was to the DRS. This might be explained as the experts not being clear on the exact meaning of "functional standards for chronological interoperability".

B2.6-8 No statements in these topics fell in the B2 category.

B3. Unsure Important and Disagree Category

The group is unsure of how these items relate to the DRS but disagree with the statements as a whole.

B3.1-3 No statements in these topics fell in the B3 category.

B3.4 DRS Focus

The DRS does not assume that digital archiving is solely a technological problem. The experts are unsure of how important this is.

B3.5 This statement topic did not fall in the B3 category.

B3.6 Metaknowledge Archive

Media metaknowledge is not rigidly defined before coming to market but the group does not see how this applies to the DRS. They do not think the DRS makes this assumption.

B3.7-8 No statements in these topics fell in the B3 category.

C1. Not Important and Agree Category

The group did not think these items directly affected the DRS but did agree on them.

C1.1 No statements in this topic fell in the C1 category.

C1.2 Media and Digital Objects

The DRS does not address what to do with the data after recovery. This is not important, as one expert stated "The DRS is concerned with data recovery not what happens to the data after recovery." In other words, let the people who wanted the data in the first place decide what they will do with it. The DRS does not address the context or order of a document in a collection and this fact is not important.

C1.3-8 No statements in these topics fell in the C1 category.

C2. Not Important and Unsure Agree Category

These items are not important and the experts cannot be sure if they agree with the statements.

C2.1-3 No statements in these topics fell in the C2 category.

C2.4 DRS Focus

The DRS may lack the archival distinction between a document and a record, but it doesn't really matter. The DRS may not address legal-related issues such as intellectual property and is not important that it does not do this. The group seems to be evenly split on the importance level of this statement. The statement might have some applicability if further clarified.

C2.5-8 No statements in these topics fell in the C2 category.

C3. Not Important and Disagree Category

These items are not important to the DRS and the group disagrees with the statements.

C3.1-3 No statements in these topics fell in the C3 category.

C3.4 DRS Focus

The DRS does not have too narrow a view of what constitutes data recovery, but this is not too important.

C3.5-8 No statements in these topics fell in the C3 category.

Appendix H: Round 3 Responses

Expert A: "I don't really have any comments on your analysis of the results"

Expert B: "I reviewed the report for round 2 and generally agree."

Expert C: "I read the Round 2 report and do not have any comments. Good luck."

Expert D: No response

Expert E: "Overall, I thought the compiled records were accurate. I think the presentation might have more bite if the points were presented in bullet form."

Expert F: "I offer the following comments on your Round 3 report, enclosed in '<>' brackets following excerpts from your report. Many of my comments are simply indications of places where I honestly could not understand the inference you were drawing from the group's responses: in some cases, this confusion seemed to stem from the form of your comments, i.e., as agreement or disagreement with statements that were often negative in form, resulting in double negatives which it was not always clear were intended.

In a few places, I have indicated my further dissent with what I interpret as the group's overall position. Feel free to ignore these comments if you like, since you have already folded my previous responses on these subjects into your group results; but I offer them as clarification of the summary results in places where I think the summary misses important arguments.

I hope this is helpful.

A1.1 Preservation and Access Environment

As young as the digital world is, we are already seeing that there is a definite need for digital archaeology.

<Make sure you define the term "digital archaeology" and use it to mean only what it really means (as it is currently being used), i.e., an approach that relies almost totally on future effort to decipher saved digital bitstreams, which is NOT a "preservation" approach in the sense that it offers no promise of being able to correctly interpret or even render material saved in this way. Whereas DRS claims not to be a preservation approach, it must still presumably serve some such approach if it is to be useful.>

A1.6 Metaknowledge Archive

The DRS can accomplish its long-term access mission because it maintains the Metaknowledge Archive.

<I think it is important to associate any statement such as this one with the caveat that it is by no means clear or proven that a MKA of the required type can be created, since there is currently no accepted or demonstrated methodology for creating the required metaknowledge, AND there is much evidence to indicate that this may be far more difficult than it sounds.>

A1.7 Software, Logical Formats and Physical Formats

Software is very important, and a concerted effort with software developers will be necessary to capture sufficient information to assist the DRS. Some files are application independent, such as .jpeg or .bmp. The "native format" is the format that the originating software used for the file and this format is important to understand.

<Note that being 'application-independent' does NOT make a file 'software-independent'. JPEG may be independent of any specific application program, but it is by no means software-independent, since it requires significant software interpretation. This is a crucial distinction, which should be brought out.>

A1.8 DRS Implementation Details

The group strongly agrees that maintaining long-term access to documents is important and that the DRS allows for that access even if no readers exist for that medium. The sentiment was not unanimous-there was one who disagreed on the DRS portion of the statement.

<My dissenting opinion here is that "access" without readability is meaningless, except from a strict digital archaeology approach. DRS (if it worked) would provide access to physical bitstreams, but this is NOT the same as "access" to a document. I feel quite strongly that this distinction gets buried in many discussions, which do not sufficiently recognize the fact that "access" to a traditional document is not at all the same thing as access to the physical bitstream of a digital document. A (poor) analogy is that of hieroglyphics prior to finding the (real) Rosetta Stone: we had "access" to the hieroglyphics, but not to the documents they represented, since we could not understand them. Furthermore, this analogy falls short of the digital case, since we would not even be able to render a digital document without being able to interpret its bitstream--once having rendered it, we would STILL face

higher-level interpretation problems such as those of knowing the language in which the document is written. A better analogy is that a digital document is written in invisible ink, which can only be made to reappear if we run appropriate software to interpret the document's logical bitstream correctly. Without this, the kind of "access" that DRS would provide would amount merely to accessing a document written in invisible ink, i.e., which would remain invisible after accessing it.>

2.7 Software, Logical Formats and Physical Formats

The format of the logical bitstream is important to the software and how the data is presented is important, but the group is unsure whether or not the DRS misses this point.

<I would argue that physical format is important ONLY if original media are the only option for access. If bitstreams are migrated to new media, then the focus should be on logical bitstreams, since there is no good reason to retain the physical formats of the original media, and the physical formats of the intermediate media (onto which the bitstreams are migrated) are of no interest to anyone. While it is possible to migrate original physical bitstream images onto new media, this is of far less relevance than capturing original logical bitstreams, which are what are required by interpreters of preserved digital documents. (Only a device controller is interested in the physical formats of original media).>

A3. Important and Disagree Category

This grouping of items was found to be important to the DRS, but the experts disagreed with the statements. This suggests a consistency in responses, because some of the statements were relatively opposite with what some of the agree statements were.

<This is unclear: I cannot figure out what it means>

A3.6 Metaknowledge Archive

<It is unclear what these statements mean: too many negatives!>

The group agrees that the MKA is important but is unsure if the MKA will be available. They do not think that the DRS makes this assumption.

<Does this mean that the group does not think that DRS assumes that the MKA will be available?>

Preserved documents are important but do not necessarily meet preservation criteria.

<What does this mean?>

The group does not think the DRS makes this assumption either.

<Which assumption?>

Media metaknowledge standards are important, but are not adhered to or valid. The group does not think the DRS makes this assumption.

<Which assumption?>

A3.7 Software, Logical Formats and Physical Formats

The group thinks that software behavior and physical format are important but that the DRS should not focus more on the software behavior than the physical format. Data re-creation is important but knowledge preservation, data recovery, and document reconstruction are not all that is needed. They also do not think that the DRS makes this assumption.

<Unclear: which assumption?>"

Expert G: "Looks OK to me. I think I said what I wanted to say in the previous set of comments. The question really is how to provide for the creation of suitable metadata/metaknowledge. Right now I see few data owners accepting responsibility for describing the data to others, and organizations like libraries don't have the resources to do this.

Thanks."

Expert H: "Thanks. This looks like a useful summary of opinions that will be helpful in guiding the DRS project - I'd say your exercise was a success.
I personally have only one argument with the summary, and it may be due to an accidental mis-statement on your part. I think that:

A3.7 Software, Logical Formats and Physical Formats

The group thinks that software behavior and physical format are important but that the DRS should not focus more on the software behavior than the physical format.
should read:

A3.7 Software, Logical Formats and Physical Formats

The group thinks that software behavior and physical format are important but that the DRS should … focus more on the software behavior than the physical format.

eg. I think the ‘not’ should be dropped.”

Bibliography

Ayton, P., William R. Ferrell and Thomas R. Stewart. "Commentaries on 'The Delphi Technique as a forecasting tool issues and analysis' by Rowe and Wright." International Journal of Forecasting, 15: 377-381 (1999).

Beagrie, Neil and Daniel Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." Final Draft of Unpublished Report. Arts and Humanities Data Service, King's College London, London SE1 7AW, 14 July 1998. n. pag. http://ahds.ac.uk/manage/framework.htm. 10 July 2000.

Bearman, David. "Reality and Chimeras in the Preservation of Electronic Records," D-Lib Magazine Volume 5 No. 4 April 1999. n. pag. http://www.dlib.org/dlib/aprill99/bearman/04bearman.html. 10 July 2000.

Brown, Bernice S. DELPHI PROCESS: A Methodology Used for the Elicitation of Opinions of Experts. The RAND Corporation, September 1968, (AD 675981).

Brown, Bernice and Olaf Helmer. Improving the Reliability of Estimates Obtained from a Consensus of Experts. The RAND Corporation, September 1964 (AD 606970).

Buckley, Chris. "Delphi: A methodology for preferences more than predictions," Library Management, Vol. 16 No. 7: 16-19 (1995).

Carlin, John W. "Records, Records Everywhere, But How Are They Going To Survive?," The Record. September, 1998.

Cox, Richard. "Access in the digital information age and the archival mission: The United States," Journal of the Society of Archivists Vol. 19 Issue 1: 25-40 (April 1998).

Dalkey, Norman C. Delphi. The RAND Corporation, October 1967 (AD 660554).

Daukantas, Patricia. "NARA's Web archive plan irks agencies," Government Computer News. 5 February 2001.

Day, Michael. "Extending metadata for digital preservation," Ariadne Issue 9. 9 May 1997. n. pag. http://www.ariadne.ac.uk/issue9/metadata. 10 July 2000.

-----. "Metadata for digital preservation: an update," Ariadne Issue 22. 21 December 1999. n. pag. http://www.ariadne.ac.uk/issue22/metadata/. 10 July 2000.

-----. "Metadata for Preservation." Report to CEDARS Access Issues Working Group, 3 August 1998. n. pag. http://www.ukoln.ac.uk/metadata/cedars/AIW01.html. 18 September 2000.

Directorate of Operational Requirements and Development Plans, Headquarters USAF. Contract F44620-67-C-0045 with The RAND Corporation. Dalkey, Norman C. "The Delphi Method: An Experimental Study of Group Opinion." Memorandum RM-5888-PR, June 1969. (AD 690498).

-----. Contract F44620-67-C-0045 with The RAND Corporation. Brown, B., S. Cochran and N. Dalkey. "The Delphi Method, II: Structure of Experiments." Memorandum RM-5957-PR, June 1969. (AD 690988).

-----. Contract F44620-67-C-0045 with The RAND Corporation. Brown, B., S. Cochran and N. Dalkey. "The Delphi Method, IV: Effect of Percentile Feedback and Feed-In of Relevant Facts." Memorandum RM-6118-PR, March 1970. (AD 702790).

Dooley, David. Social Research Methods. New Jersey: Prentice Hall, 1995.

"Dublin Core Metadata Initiative." n. pag. http://purl.oclc.org/dc. 27 September 2000.

"Dublin Core Metadata Element Set, Version 1.1: Reference Description." n. pag. http://purl.oclc.org/dc/documents/rec-dces-19990702.htm.

Encyclopedia Britannica. "Cortés, Hernán, Marqués Del Valle De Oaxaca," n. pag. http://search.eb.com/bol/topic?eu=26861&sctn=3#s_top. 26 February 2001.

Graham, Peter S., Associate University Librarian Rutgers, the State University of New Jersey. "Building The Digital Research Library: Preservation and Access At The Heart Of Scholarship." Address to Follet Lecture Series Attendees. Leicester University, United Kingdom. 19 March 1997.

Gray, Jim and Prashant Shenoy. "Rules of Thumb in Data Engineering." Technical Report MS-TR-99-100. March 2000 Revision. Proceedings of 16th International Conference on Data Engineering. 3-12.

Griffin, Ricky W. Management. Boston: Houghton Mifflin Company, 1999.

Hedstrom, Margaret. "Digital Preservation: Matching Problems, Requirements and Solutions." Essay. n. pag. http://cecssrv1.cecs.missouri.edu/NSFWorkshop/hedpp.html. 29 July 2000.

Hedstrom, Margaret and Sheon Montgomery. "Digital Preservation Needs and Requirements in RLG Member Institutions." Essay. December 1998. n. pag. http://www.rlg.org/preserv/digpres.html. 10 July 2000.

Hodge, Gail M. "Best Practices for Digital Archiving: An Information Life Cycle Approach." D-Lib Magazine Vol. 6 Issue 1. January 2000. http://www.dlib.org/dlib/january00/01hodge.html. 27 September 2000.

Intel. "Processor Hall of Fame." Essay. n. pag. http://www.intel.com/intel/museum/25anniv/hof/hof_main.htm. 2 August 2000.

-----. "What is Moore's Law?" Essay. n. pag. http://www.intel.com/intel/museum/25anniv/hof/moore.htm. 2 August 2000.

Kenney, Anne. "Digital to Microfilm Conversion: A Demonstration Project 1994-1996." Essay. 1996. n. pag. http://www.library.cornell.edu/preservation/com/comfin.html. 10 July 2000.

Kochtanek, T. R., and K. K. Hein. "Delphi study of digital libraries," Journal of Information Processing and Management Vol. 35 Issue 3: 245-254 (May 1999).

Krippendorff, Klaus. Content Analysis: An Introduction to Its Methodology. Beverly Hills CA: Sage Publications, 1980.

Lesk, Michael. "Preserving Digital Objects: Recurrent Needs and Challenges." Essay. n. pag. http://www.lesk.com/mlesk/auspres/aus.html. 10 July 2000.

Linstone, Harold and Murray Turoff. The Delphi Method: Techniques and Applications. Menlo Park CA: Addison-Wesley, 1975.

Lyman, Peter and Howard Besser. "Time & Bits Managing Digital Continuity." Essay. n. pag. http://www.longnow.com/10klibrary/TimeBitsDisc/tbpaper.html. 10 July 2000.

Lyman, Peter and Hal R. Varian. How Much Information? The University of California at Berkeley, CA. 2000.

Lynch, Clifford. "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital information," D-Lib Magazine Vol. 5 No. 9 n. pag. http://www.dlib.org/dlib/september99/09lynch.html. 10 July 2000.

MacCarn, Dave. "Toward A Universal Data Format For The Preservation Of Media." Essay. n. pag. http://info.wgbh.org/upf/papers/SMPTE_UPF_paper.html. 18 September 2000.

Matthews, William. "Webmasters flinch at 'snapshots'," Federal Computer Week. n. pag. http://fcw.com/fcw/articles/2001/0115/web-web-01-18-01.asp. 26 January 2001.

McKemmish, Sue, Adrian Cunningham and Dagmar Parer. "Metadata Mania." Records Continuum Research Group, Monash University, 1998. Essay. n. pag. http://www.sims.monash.edu.au/rcrg/publications/recordkeepingmetadata//smo1.html. 15 September 2000.

Microsoft Corporation. "Windows 2000 Professional Getting Started Guide," http://www.microsoft.com/windows2000/guide/professional/sysreq/default.asp 1999.

-----. "Office 2000 System Requirements," n. pag. http://www.microsoft.com/office/sysreq.htm. 26 February 2001.

Miller, James S. "W3C and Digital Libraries," D-Lib Magazine. November 1996. http://www.dlib.org/dlib/november96/11miller.html. 27 September 2000.

Moore, Reagan and others. "Collection-Based Persistent Digital Archives-Part 1," D-Lib Magazine Volume 6 Issue 3. March 2000. http://www.dlib.org/dlib/march00/moore/03moore-pt1.html. 27 September 2000.

National Archives of Canada. Fading Away: Strategic Options To Ensure The Protection Of And Access To Our Audio-Visual Memory. Canada: June 1995.

Norsam Technologies, Inc., "HD-ROSETTA Archival Preservation Services," n. pag. http://www.norsam.com/rosetta.html. 26 February 2001.

Pace, Andrew. "Digital Preservation: Everything New Is Old Again," Computers in Libraries: 55-58 (February 2000).

Paige, Emmett, Jr., Assistant Secretary of Defense. "Implementation of Defense Information Infrastructure Common Operating Environment Compliance." 23 May 1997.

Paliwoda, Stanley J. "Predicting the Future Using Delphi," Management Decision Vol. 21 No. 1: 31-38 (1983).

Poitras, Gilles. "QUERY - 8 inch Disks, AES Computers." Electronic Message. n. pag. http://www.fringeware.com/msg/1998/msg8292.html. 2 August 2000.

Rivette, Kevin G. and David Kline. "Discovering new value in intellectual property," Harvard Business Review: 54-66 (January-February 2000).

Robertson, Steven B. Digital Rosetta Stone: A Conceptual Model For Maintaining Long-Term Access To Digital Documents. MS Thesis, AFIT/GIR/LAR/96D-8. School of Logistics and Acquisition Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December, 1996 (AAM 8385 0 1172 0021663 0).

Rothenberg, Jeff. Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. January 1998.

Schaller, Robert. "The Origin, Nature, and Implications of 'MOORE'S LAW: The Benchmark of Progress in Semiconductor Electronics." Unpublished report to Dr. Roger Stough for Macro Policy (PUBP 801, Spring 1996). September 1996.

Smith, Abby. "Preservation in the Future Tense," Council on Library and Information Resources No. 3. May/June 1998. http://www.clir.org/pubs/issues/issues03.html. 28 June 2000.

Tait, Steven. Extensible Markup Language (XML). Unpublished report to Air Force Institute of Technology (AU), Wright-Patterson AFB OH, February 2000.

Task Force on the Archiving of Digital Information. "Preserving digital information: Report of the Task Force on Archiving and Digital information." Washington DC: Commission on Preservation and Access, 1 May 1996.

Tennant, Roy. "Digital Libraries So Much To Digitize, So Little Time and Money," Library Journal Vol. 123 Issue 13: 36-37 (August 1998).

Tersine, Richard J. and Walter E. Riggs. "The Delphi Technique - A Long-Range Planning Tool," Business Horizons Vol. 19 No. 2: 51-56 (April 1976).

United States Congress. 44 USC PUBLIC PRINTING AND DOCUMENTS. 5 January 1999.

Waters, Don. "Some Considerations on the Archiving of Digital information." Essay. n. pag. http://www.oclc.org:5046/oclc/research/links/archtf/waters_dig_arch.html. 10 July 2000.

Williams, Roy. "Data Powers of Ten." Essay. n. pag. http://www.cacr.caltech.edu/~roy/dataquan/. 15 February 2000.

Willis, Don. A Hybrid Systems Approach to Preservation of Printed Materials. Washington DC: The Commission on Preservation and Access, November 1992.

Zuzga, Brian K. Tape Archiving Using the Time Capsule File System. ME thesis. Massachusetts Institute of Technology, MA, June 1995.

Vita

Captain Don M. Kelley was born in [redacted] near Atlanta, Georgia. He graduated from Tift County High School in 1991. From there he entered Abraham Baldwin Agricultural College, and after one quarter, enrolled in Valdosta State University. While there, Captain Kelley was a cadet at ROTC Detachment 172. He earned a Bachelor of Science in Computer Science from Valdosta State. Upon graduation in 1995, his first assignment was with the Defense Information Systems Agency, Joint Staff Support Center, at the Pentagon. He worked as an Information Systems Security Officer for a number of systems.

In 1998 he was married. Shortly after the wedding, Captain Kelley's second assignment was at Wright-Patterson AFB, near Dayton, Ohio. He was assigned to the Headquarters, Materiel Systems Group, DI – Defense Information Infrastructure. While there, he worked with MSG personnel regarding security certifications and accreditations. In 1999, Captain Kelley was selected to attend the Air Force Institute of Technology. In March of 2001, he graduated with a Master of Science in Information Resource Management Degree.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
20-03-2001	Master's Thesis	Sept 2000 – Mar 2001

4. TITLE AND SUBTITLE

A DELPHI ASSESSMENT OF THE DIGITAL ROSETTA STONE MODEL

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelley, Don, M., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GIR/ENV/01M-10

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

88 WS/WEA
Attn: Captain Nathan Drummond
2049 Monahan Way, Bldg 91
Wright Patterson AFB, OH 45432-7204
(937) 255-2315 Fax 937-255-1158
DSN 785-2315

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Information that is stored digitally can only be used if it can be retrieved and interpreted. If the methods to retrieve the information are lost, it may be difficult, if not impossible, to re-create them. The knowledge to interpret the bitstream is also at risk. The Digital Rosetta Stone (DRS) Model was developed as a framework for capturing and maintaining the methods necessary to retrieve and display digital information stored on obsolete media or using obsolete software. However, this conceptual model had not yet been assessed by the community of professionals for its practical efficacy. This thesis began the assessment process by using the Delphi Method to explore the DRS with those responsible for maintaining access to digital data.

During the first round of the Delphi, the ideas expressed by the group of experts formed the basis for further discussion. Overall, the group expressed concerns about the practicality of developing the DRS, but agreed that it is an important concept that should be explored further. If found to be technologically feasible and economically desirable, the DRS could well lead to long-term solution to recovering information that would otherwise be impossible to recover.

15. SUBJECT TERMS

Delphi Techniques, Digital Systems, Data Management, Information Retrieval, Access, Preservation

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(Include area code)*
Unclassified	Unclassified	Unclassified	Unlimited	149	Associate Professor Alan R. Heminger (ENV) Alan.Heminger@afit.edu (937) 255-3636, ext 4797

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18